Les cahiers d'EDORA

Alain Pave

INRIA

UNITE DE RECHERCHE
INRIA-SOPHIA ANTIPOLIS

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél. (1) 39 63 55 11

Rapports de Recherche

N° 866

LES CAHIERS D'EDORA

Alain PAVE

JUILLET 1988

LES CAHIERS D'EDORA

RESUME

Cet ouvrage présente les résultats des premiers travaux du club EDORA surla modélisation dans les sciences biologiques. Il s'agit d'un large champ d'application de la modélisation encore à explorer, et réciproquement le modèle mathématique apparaît de plus en plus efficace dans la résolution de problèmes biologiques. Ces développements sont illustrés ici par la réalisation de logiciels et par une contribution à la réflexion méthodologique.

EDORA LETTERS

ABSTRACT

In this report, we give an account of work performed in the EDORA scientific club in the domain of modelization in biological sciences. This application field is in a large extent, still to be investigated, and convercely, mathematical model, prove more and more useful in the solution of biological problems. Both software packages and methodological contributions are reported here in, illustrating these investigations.

Avertissement et remerciements

Cet ouvrage présente les résultats des premiers travaux du Club EDORA (Equations Différentielles Ordinaires et Récurrentes Appliquées). Dans notre approche les sciences biologiques sont le domaine d'intervention privilégié. En effet, il s'agit d'un large champ d'application de la modélisation encore à explorer, et inversement le modèle mathématique apparaît de plus en plus efficace dans la résolution de problèmes biologiques. Plus généralement, on peut retenir que les domaines d'utilisation sont des repères concrets pour le développement de méthodes, en particulier celles liées à l'informatique et à la modélisation. Développements féconds illustrés ici par la réalisation d'excellents logiciels et par une contribution non négligeable à la réflexion méthodologique.

On verra aussi que dans l'activité par essence même pluridisciplinaire que développe le Club EDORA, chaque discipline y trouve son compte, chaque spécialiste un cadre de réflexion qui apparaît comme profitable, il n'y perd pas son identité. Cette dernière condition me semble nécessaire au bon fonctionnement d'une collaboration, d'un projet de ce type, au moins pour ne pas sombrer dans un amateurisme éclairé. Il faut souligner également que la qualité des relations entre personnes importe beaucoup. EDORA réunis ces conditions nécessaires. C'est sans doute pourquoi on peut être confiant dans l'avenir du Club et dans la pertinence de ses travaux, et que d'autres publications et d'autres réalisations informatiques que celles présentées ici concrétiseront encore son activité. Par exemple, on peut espérer atteindre, pour les cahiers d'Edora un rythme annuel de parution.

Outre les auteurs qui ont participé à la tâche commune dont on trouvera ici les contributions, je tiens à remercier tout particulièrement:

* P. BERNHARD et C. LOBRY pour leur initiative dans la création du club, leur participation et le soutien actif qu'ils ont apporté à ce travail,
* M.-J. PIERI et G. ALLOUCH pour la mise en forme définitive de cet ouvrage.

mars 1988 Alain PAVÉ

Sommaire

EDORA: MODÉLISATION DE SYSTEMES BIOLOGIQUES

Alain PAVÉ
Laboratoire de Biométrie et de Biologie des Populations
UA CNRS 243
Université Claude Bernard Lyon 1
69622 Villeurbanne Cedex

Ce recueil fait le point sur l'activité du groupe Edora (Equations différentielles ordinaires et récurrentes appliquées). Cette activité est centrée sur la modélisation des systèmes biologiques, en particulier, comme le prouve son nom, sur la modélisation mathématique. En introduction il nous est apparu essentiel de faire une présentation générale des travaux et des réflexions incités par cette problématique, dans le cadre de ce Club.

Ainsi, après une présentation historique et des principaux acteurs de l'aventure, nous rappelerons les principales motivations de notre travail; il s'agit en fait de la reprise d'un texte adressé en juillet 1984 au CNRS par C. Lobry et moi-même. Ensuite un sommaire commenté décrit le contenu de cet ouvrage, puis le projet est présenté ainsi que les principaux résultats et réalisations tant en ce qui concerne les aspects informatiques, en particulier les outils de développement de systèmes à bases de connaissances, que ceux relatifs au domaine d'intervention, à savoir la représentation mathématique de phénomènes biologiques. Nous concluons sur l'avenir, en remarquant que la construction de modèles mathématiques pose d'intéressants problèmes qui ont encore à être résolus; nous gardons ce fil directeur comme axe privilégié de notre activité, mais le travail effectué et la compétence acquise nous mènent à nous interroger sur des applications plus vastes en biologie des systèmes à bases de connaissances.

1. Le Club et le Projet EDORA

Depuis quelques années, plus précisément depuis un jour de novembre 1982, un groupe de biologistes, automaticiens, informaticiens français ont décidé de concrétiser une part importante de leurs activités dans la réalisation de **logiciels d'aide à la modélisation** de la dynamique de phénomènes biologiques et plus généralement de promouvoir l'**utilisation des modèles mathématiques dans les sciences de la vie** grâce justement à l'emploi d'outils informatiques adaptés. Ce groupe a pris le nom de Club Edora (Equations Différentielles Ordinaires et Récurrentes Appliquées), il a été constitué en grande partie à l'initiative de l'INRIA sur une idée de C. LOBRY, Professeur à l'université de Nice, et grâce au soutien de P. BERNHARD, Directeur de l'Unité de Recherche de Sophia-Antipolis.

En outre ce Club a décidé d'investir tout particulièrement dans un **projet "mobilisateur"** de conception d'un logiciel s'appuyant sur les acquis les plus récents en Informatique, notamment ceux relevant de l'**Intelligence Artificielle**. Ainsi s'est constitué un "noyau dur" de chercheurs consacrant une partie importante de leur temps à ce projet, il concerne divers laboratoires et institutions: les Centres de Sophia-Antipolis et Rocquencourt pour l'INRIA, le Laboratoire ARTEMIS (Grenoble), le Laboratoire de Biométrie et de Biologie des Populations (Lyon) pour le CNRS, et le Laboratoire de Biométrie du Centre de Recherche de Jouy en Josas pour l'INRA.

Au cours de nos premières réunions nous nous sommes convaincus de trois idées essentielles:

(i) les **modèles mathématiques** sont effectivement **utiles** aux biologistes, cependant leur élaboration, leur étude et leur utilisation n'ont aucun caractère d'évidence,

(ii) l'**Informatique** devrait pouvoir apporter une aide, au moins partielle, à toute tentative de modélisation; en fait, il existe déjà des logiciels dits de simulation et d'identification, ce n'est donc pas une découverte...,

(iii) par contre, ces outils apportent des solutions partielles, pas toujours satisfaisantes et parfois dangereuses à l'utilisation. Il fallait donc sortir de ces approches classiques; l'idée de recourir aux méthodes les plus actuelles a alors été retenue, en particulier celles liées à l'**Intelligence Artificielle.**

Ce dernier point paraît risqué à qui se donne des délais, des spécifications précises et des contraintes de réalisation, car en fait il faut bien admettre que malgré un optimisme souvent affiché on ne sait pas encore bien faire. Il ne nous apparaît pas exister encore de paradigme en ce domaine: en résumé l'I.A., et plus particulièrement la branche Systèmes Expert (ou plus généralement Systèmes à Bases de Connaissances), est encore largement du domaine expérimental et il n'existe pas de méthodologie de constitution de bases de connaissances. En outre, la modélisation en biologie n'était pas forcément le champ d'investigation le plus immédiat pour envisager la construction d'un logiciel d'aide à la modélisation. Nous aurions été sans doute beaucoup plus tranquilles en mécanique, en électricité, voire même en cinétique chimique ou biochimique.

C'était donc accumuler les difficultés, du moins apparemment: un champ encore imprécis, une technologie encore aux premiers balbutiements. En fait, il est apparu, soyons honnêtes *a posteriori*, que c'est probablement à l'articulation de deux difficultés que les avancées les plus intéressantes pouvaient être faites dans les deux domaines: d'un côté l'énoncé du savoir-faire de l'expert, des besoins et des contraintes qu'il exprime permet d'envisager des aspects nouveaux dans la formalisation et le traitement des connaissances, d'un autre côté la nécessaire formalisation des connaissances proposée par l'informaticien conduit l'expert (en l'occurrence le modélisateur) à s'interroger sur cette connaissance elle-même, son organisation et son traitement. On voit se dessiner alors un projet typiquement "**recherche**" de durée indéterminée. Cependant, pour éviter une dilution des efforts nous établissons nous-mêmes des **contraintes**, temporelles et matérielles, de réalisations informatiques opérationnelles et de diffusion des résultats, et c'est là, peut-être, dans la méthode qu'apparaît la dernière originalité de ce travail.

Ainsi en juin 1986 a été présentée à l'assemblée critique du Club la première maquette du logiciel Edora organisé autour du moteur d'inférence Shirka conçu à cette fin par F. Rechenmann. Comme un jeune bébé il ne disait guère plus que "âbheu, baaa, ...", à peine "papa", il avait tendance à se laisser aller aussi, mais ses réactions encourageantes nous ont fait penser que le bébé n'était pas un monstre constitution-nellement débile et qu'il pourrait grandir, apprendre et devenir un jour un jeune logiciel adulte BCBG que nous espérons, comme tous parents, beau et intelligent. Mais nous sommes conscients comme devraient l'être tous les parents que la partie génétique ne fait que proposer un moule et qu'il est de notre responsabilité de l'améliorer, et de faire son éducation.

C'est aussi pourquoi dans un proche avenir on trouvera sur quelques micro-ordinateurs des réalisations comme DYNAMAC qui révolutionne la manière d'approcher un système différentiel ordinaire, et dont l'objectif était, comme CROISSANCE, autre avatar du projet, de tester certaines idées d'interaction avec l'utilisateur.

Enfin, il nous est apparu très tôt que l'histoire d'Edora ne doit pas se disséminer dans des éditions aussi diverses que polyglottes, et que s'il est bien et nécessaire que nos travaux apparaissent dans la littérature nationale ou internationale, il est au moins aussi bon que les travaux, réflexions, compilations et synthèses auxquels ce projet a conduit soient récapitulés dans une publication par essence même apériodique: **les Cahiers d'Edora.**

2. Principales motivations

Les tentatives de représentations mathématiques de phénomènes biologiques ne sont pas nouvelles (on retiendra par exemple que dès le 13ème siècle, Fibonnacci prenait comme support de réflexion un problème qu'on classerait maintenant dans le champ de la dynamique des populations), cependant la pertinence des "modèles" obtenus pour l'analyse, voire le contrôle, de ces phénomènes, ou l'aide à l'expérimentation, n'est apparue de façon claire que récemment. Il a fallu d'une part l'invention de méthodes et de moyens d'étude qualitative et quantitative, d'autre part une évolution suffisante des connaissances et des problématiques biologiques d'où ressort plus nettement le rôle possible des outils mathématiques disponibles (par exemple, on est encore mieux armé pour formaliser les aspects dynamiques et fonctionnels que les aspects descriptifs et structuraux, alors que ces derniers ont en premier retenu l'attention des biologistes).

Cependant, au stade actuel d'évolution de l'art, on peut faire les remarques suivantes:

- du côté méthodologique, des compétences diverses doivent être réunies pour aborder un problème de modélisation (mathématiques, informatiques, statistiques, automatiques..., et évidemment biologiques); elles sont difficilement maîtrisables par une seule et même personne, et même à rassembler au sein d'une équipe;

- du point de vue biologique l'énoncé du problème et l'approche expérimentale doivent tenir compte le plus précisément possible de l'approche "modélisation". C'est d'ailleurs une remarque plus générale qui vaut pour d'autres outils: leur existence et les contraintes liées à leur utilisation doivent être considérées;

- la demande du secteur expérimental, très faible il y a quelques années, devient de plus en plus importante. Par exemple, au Laboratoire de Biométrie de l'UCB (Lyon 1), pour les seules activités modélisation, quelques actions ponctuelles ont été tentées entre 1970 et 1980, correspondant à l'activité d'un ou deux chercheurs. Actuellement, et bien que l'équipe "modélisation" se soit considérablement étoffée (4 personnes en poste, et en moyenne 6 thésards), il n'est plus possible de faire face à la demande.

Pour répondre à ces besoins, la communauté scientifique française a proposé quelques solutions:

- une association comme l'AMTB (Association pour le développement des Méthodes Théoriques en Biologie) a un programme régulier de cours de formation permanente de chercheurs, la SFB (Société Française de Biométrie, correspondante de la Société Internationale de Biométrie) organise une réunion annuelle dans le cadre des journées de Statistique de l'Association des Statisticiens Universitaires, elle a contribué à l'édition d'ouvrages originaux (Biométrie et Ecologie, Biométrie en biologie cellulaire, ...);

- beaucoup de laboratoires se dotent d'une compétence informatique (généralement autodidacte) permettant de pallier les besoins les plus pressants (essentiellement la saisie des données et leur traitement statistique);

- quelques laboratoires, encore trop peu nombreux en France (Département de Biométrie de l'INRA et certains laboratoires universitaires) cherchent à intégrer les diverses compétences, mais ne peuvent faire face à tous les besoins;

- enfin, la collaboration entre biologistes et mathématiciens est de plus en plus encouragée (par exemple, par des ATP (Actions Thématiques Programmées)), mais connaît une limitation de principe: les chercheurs en mathématiques refusent de pratiquer une activité de service; c'est normal car leur fonction essentielle est de développer leur propre champ disciplinaire et on ne peut pas leur reprocher de le faire mal (l'école

mathématique française est encore une des meilleures du monde et il faut qu'elle le reste). En outre les questions encore posées en biologie ne relèvent encore que très rarement d'une mathématique pointue où le mathématicien pourrait trouver un intérêt. Inversement, pour classique qu'elle soit, cette mathématique est peu familière au biologiste. En attendant la formation et le recrutement significatifs (???) de chercheurs et d'ingénieurs Biométriciens ou plus généralement en "Mathématiques Appliquées à la Biologie", le problème reste entier. Signalons cependant que des tentatives sont menées dans certaines Universités (par exemple, des U.V. spécialisées de 2ème cycle, comme le MAB (Mathématiques Appliquées à la Biologie) à Lyon ou celle réservée aux étudiants de la Maîtrise de Génie Biologique, DEA et formations doctorales de Biomathématiques à Paris et à Lyon).

Si bien qu'il nous est apparu que des propositions devaient être faites pour surmonter ces difficultés. Outre des actions de formation nécessaires, une réponse possible consistait à regrouper des compétences autour d'un projet commun de logiciel d'aide intelligente à la modélisation. Ainsi faisons-nous fonctionner une double structure: celle d'un **Club** assurant le meilleur environnement scientifique possible, et celle d'un **Projet** regroupant les développeurs du logiciel; Club et Projet ayant adopté le même sigle: **Edora** . En plus du produit fini, il apparaît très nettement aujourd'hui que même pendant la phase de développement la conjugaison des efforts peut contribuer à la fois à l'extension de la méthodologie de la modélisation en biologie et au développement d'outils informatiques originaux; comme on l'a signalé plus haut, c'est peut-être la qualité essentielle d'un bon projet.

3. Premiers résultats

Dans un premier temps, il était nécessaire de délimiter le champ d'intervention. Nous avons choisi de nous attaquer aux problèmes de modélisation de la dynamique de phénomènes biologiques: problèmes actuels qui correspondent à la fois aux compétences des membres du Club et à une évolution significative de la demande en Biologie. D'autre aspects auraient pu être choisis, par exemple l'analyse des données ou la statistique plus classique, dans les deux cas d'autres équipes abordent le problème (par exemple, on pourra se référer au récent ouvrage de **Gale:** Artificial Intelligence & Statistics, *Addison-Wesley*, 1986), alors qu'en modélisation des phénomènes biologiques à notre connaissance peu de travaux sont actuellement développés comme en témoigne l'ouvrage collectif publié par **Kherkoffs et al** (A.I. Applied to Simulation, *SCS Publ.*, 1986) sur les aspects modélisation et simulation.

3.1. Etat des lieux et réfexions préliminaires

Il existe de nombreux programmes plus ou moins sophistiqués de **simulation**, cependant, aucun, à notre connaissance, n'intègre l'ensemble des outils et des compétences permettant une utilisation efficace et surtout raisonnable de ceux-ci. Aussi avons-nous lancé une réflexion, doublée de réalisations, pour vérifier nos idées, qui nous ont permis d'étudier la faisabilité d'un système intégré. Ce système a comme caractéristiques d'aider l'utilisateur non seulement sur les plans classiques du calcul (numérique et formel), de la gestion des données et des modèles, mais aussi sur ceux de l'étude qualitative, de la sélection argumentée des algorithmes à mettre en oeuvre, voire même sur celui du choix ou de la construction d'un modèle en fonction des données, du problème, de l'objet biologique concerné et des objectifs de la modélisation.

Ainsi, outre les outils classiques (élaborés, choisis, adaptés et critiqués par des spécialistes), ce logiciel doit contenir la connaissance liée:

- aux méthodes offertes, aux objets manipulés (formels, mathématiques et données biologiques),

- au domaine d'application (dynamique des systèmes biologiques),

- à la forme des connaissances manipulée (e.g., connaissances sous forme déclarative ou procédurale)

On tient compte du fait qu'un objet, par exemple un modèle, peut être regardé de différentes façons, en l'occurence d'un point de vue purement formel, d'un point de vue mathématique, d'un point de vue procédural (quelles sont les méthodes qui lui sont applicables), ou encore d'un point de vue biologique; de même il faut considérer que l'utilisation d'un algorithme dépend du contexte (type de modèle, qualité des données...).

Enfin il est clair que doivent être prévues des interfaces de dialogue commodes qui rendent le logiciel efficace, simple et agréable à utiliser.

3.2. Travaux effectués

Pour tester ces idées, et plutôt que définir une progression en série longue et limitée par la réalisation de certaines tâches, nous avons mené un ensemble de travaux en parallèle:

- Réflexions et compilations théoriques sur
  . les modèles et la modélisation en biologie, le présent recueil fait le point sur ce sujet (cf.§ 4),
  . la représentation et la manipulation des connaissances (déclarative et procédurale), leur adéquation au problème posé (cf. § 5).

- Logiciels:
  . développement d'un système de gestion de bases de connaissances en représentation centrée-objet: SHIRKA (cf. § 6),
  . réalisation de maquettes et de programmes spécifiques: notamment EDORA V.1, DYNAMAC, CROISSANCE, MODEL et SIMUL (cf. § 7).

3.3. Choix techniques

Les choix qui ont été faits tiennent largement compte de l'évolution des matériels, des logiciels et des habitudes liées à cette évolution:

- nous avons retenu le concept de poste de travail illustré par des matériels de type Sun, Apollo, SPS 7 ou à un niveau micro par le Macintosh d'Apple (notamment le Mac II), comme systèmes cibles pour une telle application;

- cependant les environnements de développement sont divers: DPS-8 sous Multics, SPS-7 et Sun sous Unix, Vax sous VMS, et surtout Macintosh. Ils correspondent à la diversité des systèmes accessibles dans les laboratoires concernés, et au fait qu'Edora n'a pas, en tant que tel, les moyens permettant un équipement spécifique au projet;

- les aspects I.A. sont développés en Le_Lisp de l'INRIA, disponibles sur les systèmes cités;

- les logiciels et programmes plus classiques sont écrits en Fortran 77, en Pascal et en C.

3.4. Premières conclusions

Dès à présent, on peut noter

- d'une part que les travaux liés à la conception de ce type de système informatique traitant de connaissances sous forme déclarative ou procédurale, et relevant de différents domaines (mathématique, statistique, biologique,...), conduisent à des résultats suffisamment généraux pour être utilisables dans des champs d'applications autres que celui de la modélisation en biologie;

- d'autre part que la nécessité d'une formalisation et d'une organisation de la connaissance n'est pas sans influence sur le champ d'application lui-même (c'est ce qui est notamment apparu pour les modèles "classiques" de la biologie des populations).

4. Modèles et modélisation en biologie - Les Cahiers d'Edora

L'essentiel des contributions initiales fait l'objet de la présente publication d'Edora: l*es cahiers d'Edora*. En effet il est apparu intéressant, à part les aspects proprement logiciels, que les idées développées dans le cadre du Club ou du Projet apparaissent non seulement dans la littérature traditionnelle, mais, étant donnée leur originalité due à une certaine unité de pensée, qu'elles puissent être exposées dans des publications plus ou moins périodiques. C'est ainsi que ce premier numéro est consacré à une mise au point méthodologique présentant essentiellement divers aspects de la modélisation en biologie. La plupart de ces travaux ont, par ailleurs, fait l'objet de publications dans des revues ou dans des actes de colloques.

Les problèmes suivants ont été abordés:

- *Construction et interprétation de modèles mathématiques en biologie* (A. Chéruy, S. Gentil, F. Houllier, A. Pavé): les contributions correspondantes proposent une méthodologie et des outils de choix ou de construction, prenant en considération le type d'objet biologique, la nature des données, et les objectifs ou la problématique posés; les idées venues de l'automatique ont été déterminantes (A. Chéruy, S. Gentil, *op. cit.*). En outre, on s'est intéressé à certains modèles "classiques" (entendre modèles différentiels) de dynamique des populations et de croissance individuelle d'organes ou d'organismes dans l'optique d'organisation d'une base de connaissance concernant ces modèles (A. Pavé, F. Houllier, *op. cit.*). A cette occasion des problèmes relatifs aux niveaux de connaissances et à leurs relations sont soulevés (niveau phénoménologique ou superficiel, niveau explicatif ou profond). Par exemple, un modèle de croissance est sensé décrire la croissance d'une population ou d'un organisme, on pourrait ainsi dire qu'une croissance est du type logistique ou de Gompertz (niveau du phénomène observé); le couplage avec un ou des schéma(s) fonctionnel(s) peut conduire à une interprétation plus mécaniste: croissance sur un milieu limité en ressources, ou croissance régulée par un facteur de croissance (niveau explicatif).

- *Identification et validation*: les points de vue de l'automaticien et du statisticien sont discutés (S. Gentil, A. Messéan). Les principales méthodes d'identification sont référencées (méthodes des moindres carrés, maximum de vraisemblance). En fonction du modèle retenu pour l'erreur de mesure, les propriétés des estimateurs des paramètres sont présentées (précision, corrélation). La plupart des modèles référencés sont non-linéaires en fonction des paramètres; les méthodes d'identification sont donc itératives. Se pose ainsi le problème des valeurs initiales, une note technique présente une solution pour certains modèles fréquemment rencontrés (F. Houllier et A. Pavé).

- *Aspects mathématiques, analyse qualitative*: le type de modèle qui nous intéresse s'exprime le plus souvent sous forme d'équations différentielles, il est donc important d'avancer dans la connaissance de ces objets, ainsi L. Baratchard, E. Benoit et J. Grimm proposent des résultats concernant notamment les "systèmes à second membres quadratiques".

- *Etudes particulières*: il est souvent intéressant d'illustrer, voire même de détecter, les problèmes posés par l'utilisation d'un modèle mathématique à travers la critique de son usage dans la littérature, c'est ainsi que J.L. Gouzé et A. Sciandra prennent comme exemple le modèle exponentiel; ils le comparent notamment à d'autres modèles et montrent que ces derniers s'ajustent au moins aussi bien aux données expérimentales: on en tire donc que le critère de "bon" ajustement ne suffit pas toujours pour valider un modèle notamment si celui-ci est utilisé à des fins explicatives, mais que ce critère doit être complété par des considérations biologiques.

- *Aspects informatiques*: les principales idées de mise en oeuvre informatique, autour de la notion de système à base de connaissance et du concept de poste de travail, sont présentées par B. Rousseau et F. Rechenmann qui proposent solutions et réalisations après un examen des outils existants et la discussion de leurs imperfections; de son côté C. Pierret-Golbreich présente quelques aspects formels sur l'intégration des objets symboliques dans Edora .

Enfin des travaux relevant de l'activité du Club ont été exposés au cours de séminaires et publiés par ailleurs, certains d'entre eux seront repris sous une forme adaptée dans le prochain numéro des Cahiers d'Edora.

5. Représentation des connaissances et mécanismes de raisonnement associés

La première publication soulevant le problème "système expert et modélisation" date de 1983 (de Swaan Arons, *Mathematics and Computers in Simulation,* 1983), l'auteur proposait, à travers un exemple, d'utiliser la représentation des connaissances sous forme de **règles de production.** Dès le début de notre projet il est apparu que cette représentation avait pour ce type d'application un certain nombre de défauts, en particulier de diluer la notion d'objet à travers ses propriétés dans la base de connaissances ou encore de ne pas traiter, du moins de façon satisfaisante, des relations avec la connaissance procédurale. Aussi F. Rechenmann a-t-il proposé une autre approche, à l'époque peu développée, se fondant sur la **représentation centrée-objet** (Rechenmann, 1984 [20], 1985 [22], Bensaïd et al, 1985 [23]), l'adéquation avec notre problème fut discuté notamment par A. Pavé et F. Rechenmann (Pavé et Rechenmann, 1986 [5]). Enfin, on pourra trouver une présentation plus complète dans la notice de présentation de SHIRKA (F. Rechenmann, 1987, [32]).

5.1. La Représentation Centrée-Objet

L'élément de base de cette représentation est un schéma (notion proche de celle de Frame proposée par Minsky (Minsky,In *"Psychology of Computer Vision"* ed. P.H. Winston, MacGrawHill, 1975)); ce schéma peut représenter:

- une classe d'objets dont les propriétés communes sont spécifiées par des attributs, on parle alors de schéma de classe (les objets peuvent être des modèles, des méthodes, des objets "biologiques", des descriptions de problèmes typiques, ...),

- des objets particuliers des classes référencées (spécimens, réalisations ou instances suivant le vocabulaire choisi). Ainsi, à un modèle auquel on aura affecté des

valeurs numériques particulières pour les paramètres (par exemple, après un ajustement) correspondra une **instance** de ce modèle.

Aux attributs sont attachés les notions de type (type simple prédéfini ou complexe défini par un autre schéma de classe), de domaine de valeur (l'ensemble des valeurs que cet attribut peut prendre) et de valeur (la valeur de cet attribut ou la façon de l'obtenir); les items correspondant aux diverses fonctionnalités au niveau des attributs sont des facettes (on trouvera par exemple les facettes $un, pour le type, $domaine, $valeur...). Alors que la définition des schémas de classes et des attributs est sous la responsabilité du concepteur de la base de connaissance, les facettes sont prédéfinies. La forme générale est:

nom du schéma

attribut 1	$facette 1 1 ...
	$facette 1 2...
	...
	$facette 1 p_1...
...	
attribut n	$facette n 1 ...
	$facette n 2...
	...
	$facette n p_n...

Les schémas de classe sont organisés en hiérarchies, les schémas les plus généraux se trouvant au sommet de la hiérarchie correspondante, les plus spécialisés comme feuilles de l'arbre correspondant.

5.2. Les mécanismes d'inférence et de raisonnement

Sur cette représentation ont été définis plusieurs mécanismes d'exploitation. L'un des principes de base consiste à créer des instances incomplètes et à essayer d'obtenir la valeur des attributs manquants par les mécanismes élémentaires suivants:

- l'**instanciation** correspondant à la création d'une instance.

- l'**héritage**: les schémas de classes qui sont des spécialisations de schémas plus généraux héritent des attributs de ces derniers et de leurs spécifications correspondantes (le domaine peut cependant être redéfini dans le sens d'une restriction). Un schéma de classe peut hériter de hiérarchies parallèles (héritage multiple).

- l'**attachement procédural** permet de calculer, si nécessaire, la valeur d'un attribut. Cette possibilité permet notamment d'envisager des liens avec des programmes de calcul numérique, voire même de calcul formel.

- **valeur par défaut** permet d'affecter une valeur par défaut à un attribut, cette valeur sera valable pour tous les schémas de classes plus spécifiques, sauf indication contraire. Cette spécification permet de faire du raisonnement suivant une logique non-monotone, elle est très utile dans "la vie courante".

- le **filtrage:** ce mécanisme permet de trouver dans la base des **instances** d'un schéma de classe suivant la valeur d'un (ou de plusieurs) attribut (s). De façon à pouvoir construire des filtres pour les schémas de classe eux-mêmes ceux-ci peuvent être considérés comme instances de schémas plus généraux: les méta-schémas.

- la **classification** (ou encore **identification** d'un objet dans une classification): une instance incomplète est créée au plus haut niveau d'une hiérarchie, il s'agit alors d'essayer de l'accrocher à un (ou plusieurs) schémas de classe le plus bas possible dans la

hiérarchie (i.e. de trouver le, ou les, schémas les plus spécialisés correspondant à une description).

Des **mécanismes d'explication** sont disponibles, citons tout particulièrement ceux attachés à la classification qui permettent d'obtenir les raisons d'un succès ou d'un échec.

5.3. L'adéquation au problème posé

Comme signalé plus haut, l'adéquation au problème de la modélisation a été discutée dans plusieurs contributions notamment: Rechenmann, 1985 [21] - Pavé et Rechnemann, 1986 [5]- Pavé, 1987 [8]. On retiendra:

- la notion d'objet, elle-même, qui permet de représenter sous une forme unique et particulièrement agréable la connaissance liées à diverses entités (objets mathématiques, objets biologiques, méthodes, situations expérimentales, données, classes de problèmes typiques, ...);

- l'organisation de la connaissance qui conduit à l'établissement de liens entre divers objets, et qui permet divers **points de vue** (par exemple, un objet biologique peut être vu dans l'optique du systématicien ou de l'évolutionniste, c'est à dire dans une classification systématique ou phylogénétique, ou encore comme élément d'une population intervenant dans un système naturel: système écologique ou agronomique, c'est à dire dans l'optique de l'écologiste ou de l'agronome);

- le mélange possible de connaissances sous formes déclarative et procédurale.

5.4. SHIRKA: un outil de développement de systèmes à bases de connaissances en Représentation Centrée Objet.

La mise en oeuvre des idées exposées ci-dessus a conduit à la réalisation d'un produit original, SHIRKA, qui présente plusieurs avantages par rapport à d'autres systèmes utilisant la représentation centrée-objet:

- Le modèle utilisé par SHIRKA est totalement uniforme. Tout schéma est lui-même une instance d'un schéma de classe de plus haut niveau (méta-schéma). De même tout attribut, facette et valeur de facette est une instance de schéma. Il en résulte un code particulièrement concis, facilement extensible et adaptable.

- L'inférence des valeurs d'attributs indéterminés ne fait pas appel à des règles, mais à des filtres décrits par des schémas. Il n'y a donc pas combinaison de plusieurs représentations de connaissances (comme dans les systèmes mélangeant règles et objets), avec les problèmes que de telles combinaisons induisent.

- Plusieurs mécanismes d'inférence sont disponibles, tous ceux décrits ci-dessus, à savoir: héritage, attachement procédural, filtrage, valeur par défaut, instanciation et classification L'attachement procédural, qui consiste à associer une ou plusieurs procédures à un attribut, utilise des descriptions externes, en termes de schémas, des procédures employées. Leur appel est donc entièrement contrôlé par SHIRKA: obtention des valeurs des paramètres et vérification de leur validité.

- Les attributs peuvent être mono ou multi-valués. Il est donc possible de traiter explicitement les listes de valeurs, globalement ou élément par élément.

- SHIRKA offre les moyens de gérer la cohérence entre les instances par l'intermédiaire de procédures attachées aux attributs et activées lors de certaines manipulations sur leurs valeurs, telles que l'ajout, la modification et la suppression. Un schéma de classe étant une instance d'une autre (meta-)classe, cette gestion de cohérence peut être étendue aux classes elles-mêmes.

- Shirka est bien adapté au développement d'applications pour lesquelles doivent coexister une base d'objets et de données, une base de méthodes algorithmiques et une base de connaissance.

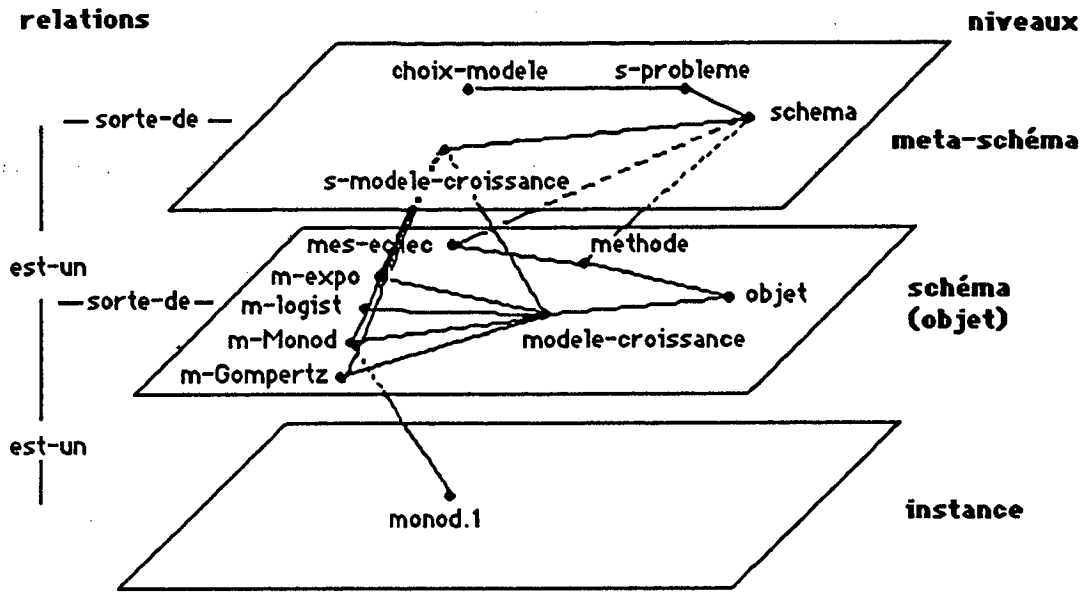


Figure 1 - Organisation de la connaissance dans SHIRKA, exemple de modèles élémentaires de croissance. La connaissance peut être organisée à trois niveaux, les schémas du niveau objet (schémas de classe) peuvent être considérés comme des instances de schémas écrits au niveau méta-schéma et donc manipulables comme des instances "ordinaires" (i.e. des réalisations de schémas de classes), en particulier pour le mécanisme de filtrage. Les relations verticales sont du type est-un alors que les relations horizontales sont du type sorte-de. [8][32].

SHIRKA est écrit en Le_Lisp; on peut l'utiliser sur tout matériel supportant Le_Lisp et son terminal virtuel et disposant de plus de 700K en mémoire centrale. Le système Shirka comporte, outre son moteur d'inférence: un langage de commande auto-documenté, un module d'explication, une interface assurant toutes les manipulations sur les instances, y compris l'interrogation des valeurs d'attributs, un éditeur qui connait la structure des schémas de classe et en facilite la création et la modification.

Enfin la connaissance est organisée sur trois plans (figure 1): les deux premiers pouvant contenir des hiérarchies de classes (méta-schémas de classes et schémas de classes), le troisième des instances de schémas de classes, les schémas de classes peuvent être considérés comme des instances de méta-schémas.

6. Conception générale du système Edora, maquettes et logiciels spécifiques

Edora est un système intégrant plusieurs types d'expertises (biologique, méthodologique, mathématique, numérique et statistique) qui cohabitent dans ce même système. Le noyau central du système est constitué par SHIRKA, l'architecture générale est du type de celle décrite par la figure 2.

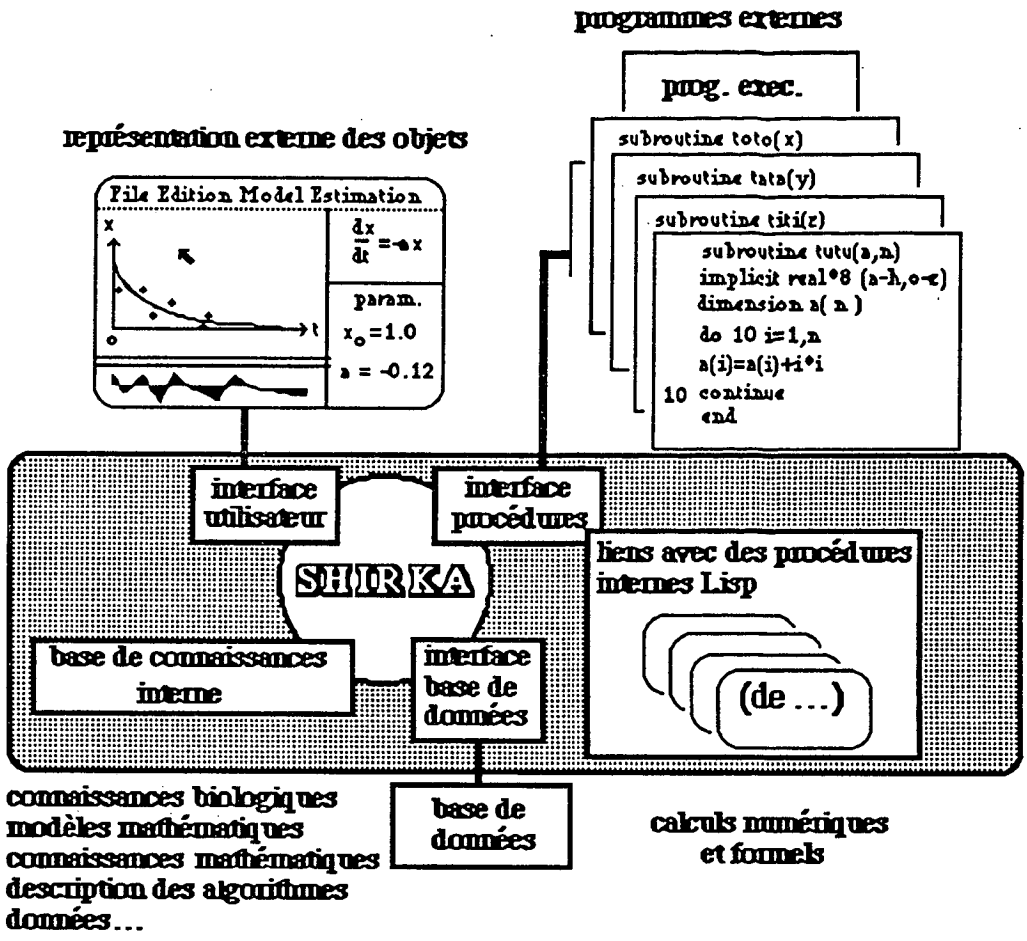


Figure 2 - Schéma général du système Edora, on notera que la représentation uniforme permet d'empaqueter l'ensemble des éléments nécessaires (connaissances déclaratives, méthodes, description d'interfaces). Il est probable que ce type d'architecture préfigure celles des futurs systèmes d'aide à la modélisation.

6.1. EDORA V.1

Une première maquette a été présentée au forum Edora de Juillet 1986, elle a été reprise sous une forme voisine et présentée au 40ème anniversaire de l'INRA en Octobre 1986.

Ce système permet de choisir un (ou plusieurs) modèle(s) de croissance dans une base de modèles en fonction de la description géométrique de la forme des données expérimentales, puis d'estimer les paramètres du modèle choisi à partir de ces données. Le système tient un cahier de laboratoire et peut résumer l'ensemble des essais effectués. Le dialogue se fait via une interface utilisateur où l'aspect graphique est prédominant.

L'implantation a été faite sous Multics pour des raisons de facilité.

6.2. DYNAMAC

Ce programme permet l'étude graphique interactive des systèmes différentiels ou récurrents sur MacIntosh. L'utilisateur commence par décrire ses équations à l'aide d'un mini-langage type Pascal. L'étude du système repose ensuite sur le tracé graphique de ses solutions dans un plan de l'espace de phase.

Le mécanisme de base consiste à choisir des conditions initiales en se positionnant dans le plan choisi à l'aide de la souris, puis à cliquer pour lancer l'intégration. De nombreuses options complémentaires permettent le choix d'une méthode d'intégration, le tracé du champ de vecteurs et le tracé des isoclines (dans le cas de systèmes différentiels plans), la représentation des courbes d'évolution en fonction du temps ainsi que la localisation et la détermination des points d'équilibre. Enfin des fonctionnalités d'édition permettent une présentation des résultats et la création de fichiers compatibles avec un logiciel de dessin .

6.3. CROISSANCE

Ce programme est conçu pour l'identification des modèles de croissance à une variable d'état. Outre les calculs classiques (estimation initiale, puis estimation par la méthode de Gauss-Marquardt), il permet une "identification à main levée" qui illustre bien la notion de travail "papier-crayon". Chacune des courbes de croissance est caractérisée par des points de contrôle dépendant des paramètres et dont les emplacements dans le plan variable-temps définissent entièrement la forme de la courbe. Par exemple, pour le modèle logistique, ces points de contrôle sont le point d'inflexion, l'ordonnée à l'origine et les asymptotes inférieure et supérieure. Pour ajuster la courbe sur les points expérimentaux, il suffit de déplacer l'un des points de contrôle à l'aide de la souris. La courbe se déforme alors, et la modification de forme est immédiatement répercutée sur les valeurs des paramètres.

Plusieurs modèles peuvent être choisis (logistique, Gompertz, logistique généralisé, monomoléculaire, ...), enfin des informations peuvent être obtenues sur la qualité des estimations des paramètres et de l'ajustement.

6.4. SIMUL et MODEL

Le programme SIMUL permet aussi la simulation (intégration numérique) d'un système différentiel en dimension n, c'est en fait un sous ensemble de MODEL qui permet de suivre une approche complète, mais classique, de modélisation numérique: intégration, identification, lissage des données, étude de sensibilité, recherche de points d'équilibre. Le logiciel graphique utilisé pour le développement est GKS (norme internationale). Ces programmes constituent des outils de base, encore imparfaits, qui seront interfacés avec Edora. Enfin ils permettent d'étudier les réactions des utilisateurs biologistes pendant leur utilisation.

Outre leurs intérêts propres ces réalisations ont permis de tester les idées que nous avions au départ et la pertinence de l'approche tant en ce qui concerne les aspects I.A. que ceux liés au concept de poste de travail bien illustré par ces trois dernières réalisations.

7. Conclusion et perspectives

L'analyse de la situation montre que:

- ce projet a cristallisé l'effort d'un groupe de personnes appartenant à trois institutions de recherche publique: INRIA, INRA et CNRS (par le biais de membres d'Unités Associées);

- des résultats tangibles et concrets ont été obtenus:
  . deux logiciels interactifs (développés par B. Rousseau) pour l'étude de problèmes spécifiques (Dynamac et Croissance) sont opérationnels,
  . un outil de développement de **systèmes à bases de connaissances** (SHIRKA) a été conçu essentiellement sur la base des questions posées par la modélisation; cet outil semble tout à fait intéressant, et susceptible d'un champ d'application plus large que celui auquel il était destiné au départ;

- deux maquettes voisines dans leur conception ont été présentées, l'une à l'INRIA en juillet 86, l'autre à l'INRA en octobre 86 (cf. § 6.1). Ces maquettes permettaient de choisir un modèle de croissance en fonction de critères de forme, d'estimer les paramètres du (ou des) modèle(s) retenu(s) à partir de données expérimentales, de tracer divers graphes, et de mémoriser les différents essais;

- deux autres maquettes présentées en 87 intégrent divers aspects formels (possibilité d'entrer un modèle nouveau via un langage d'entrée, reconnaissance de modèles, calculs formels élémentaires), et un embryon de connaissance biologique permettant de guider le choix d'un modèle de croissance;

- déjà ce projet a conduit à une trentaine de publications ou communications publiées (dans des congrès internationaux disposant d'une procédure de sélection).

En outre ont été clairement établis et mis en évidence:

- la pertinence de la représentation des connaissances choisie (représentation de type centrée objet) et des mécanismes d'inférence développés sur cette représentation;

- la profonde interaction entre informaticiens développant les outils centraux et les autres chercheurs plus orientés vers l'utilisation de ces outils, ce qui a conduit notamment les premiers à affiner progressivement la représentation des connaissances et à proposer des méthodes de raisonnement originales;

- la modélisation en biologie est un bon support de réflexion pour acquérir un savoir faire, en particulier pour étudier les problèmes de formalisation des connaissances de natures différentes afin de constituer des bases opérationnelles;

- la nécessité de formaliser la connaissance a d'importantes retombées sur le secteur d'application (bien que très spécifique et limité, le cas des modèles de la dynamique des populations est de ce point de vue très instructif); dans l'immédiat on peut raisonnablement penser qu'il s'agira là des conséquences les plus significatives de notre approche "système à base de connaissances";

- la diversification des demandes pour de tels systèmes dans les secteurs biologique, écologique, agronomique et biomédical ainsi que pour les aspects biométriques autres que la modélisation en termes mathématiques.

En fait, on remarquera plus généralement que les formalismes proposés dans les systèmes à bases de connaissances (règles de production, représentation centrée objet, ...) associés à des mécanismes de manipulation ("raisonnements"), peuvent être comparés à d'autres formalismes, en particulier au formalisme mathématique. En ce sens, ils permettent de proposer des modèles autres, c'est-à-dire de représenter des connaissances plus vastes que le strict formalisme mathématique, et de les manipuler. On peut alors penser que la démarche mise au point pour la modélisation mathématique pourra sans doute être reprise, au moins dans ses grandes lignes, pour la construction de bases de connaissances. La remarque avait d'ailleurs été déjà faite pour les bases de données (Gouy). Enfin, on peut noter, que le modèle mathématique, ses propriétes qualitatives, certaines manipulations formelles, numériques et graphiques, est parfaitement intégrable dans ce cadre formel plus général (Edora en est déjà un exemple). On voit donc arriver des systèmes à bases de connaissances intégrant divers niveaux de

formalisation, et par là même moins réducteurs que le modèle mathématique mais également plus rigoureux que l'expression verbale.

Problèmes à résoudre pour le développement d'un système d'aide à la modélisation mathématique.

Brièvement, nous avons référencé les points suivants:

(i) extension de la base de modèles,
(ii) entrée des modèles: différents langages (formulation mathématique, schémas fonctionnels divers, langages graphiques),
(iii) reconnaissance de modèles (à une paramétrisation près),
(iv) base de méthodes numériques pour l'intégration et l'identification,
(v) manipulations formelles, en particulier la définition et l'implantation des fonctionnalités élémentaires, puis la liaison avec des logiciels de calcul formel (Reduce, Macsyma),
(vi) représentation des connaissances biologiques,
(vii) gestion parallèle des connaissances,
(viii) sorties graphiques, jauges,...
(ix) formalisation de la démarche: étapes de la construction ou du choix d'un modèle, interprétation, étude et utilisation.

Ce dernier aspect pose un certain nombre de problèmes de fonds:
. préciser la notion de point de vue (celui de biologiste, du biométricien, du mathématicien...),
. pour la construction il apparaît nécessaire d'introduire **l'élaboration incrémentale** d'un modèle (correspondant à l'assemblage de processus élémentaires), de même pour l'interprétation avec **l'analyse décrémentale** (i.e. la décomposition en éléments plus simples correspondant à des processus élémentaires). Donc il est souhaitable de spécifier (au moins) deux niveaux de connaissances: le niveau phénoménologique (celui de l'observation), et le niveau des processus (aspect explicatif ou connaissance "profonde"). Ainsi en prenant l'exemple de la croissance modélisée par le modèle de Gompertz (Pavé et al 1986) [6]. On peut dire que la croissance de certains animaux est bien **décrite** par ce modèle (on dira par exemple que la croissance est du type Gompertz). Par contre quand on interprète le modèle en utilisant un schéma fonctionnel qui fait intervenir un facteur de croissance, on aboutit à un niveau explicatif (plus "profond) car on peut interpréter la croissance comme résultant d'un processus de croissance élémentaire catalysé par un facteur de croissance, ce même facteur étant lui-même dégradé dans un processus spontané de type exponentiel.

Le phénomène correspond en quelque sorte à l'observation de la sortie d'une boîte noire (connaissance phénoménologique, ou superficielle), l'analyse par décomposition en processus revient à ouvrir la boîte ou à faire des hypothèses sur des éléments plus fins qui la constituent (remarquer que la démarche est récursive car les processus qu'on considère comme élémentaires peuvent alors constituer eux-mêmes des boîtes noires; cependant il semble raisonnable de s'en tenir à **deux niveaux**). On remarquera qu'une bonne maîtrise de la théorie des systèmes est nécessaire pour bien comprendre ces problèmes.

Sur le fond, l'étude des relations entre niveaux de connaissances semble possible dans Edora car nous disposons de formalismes probablement bien adaptés.

Autres questions

Il s'agit d'une liste de questions, formulées explicitement dans divers laboratoires, dont on peut penser qu'elles relèvent d'une modélisation autre que mathématique, ou n'intégrant que localement de tels modèles, modélisation fondée sur la notion de "Système à Bases de Connaissances" et utilisant plus particulièrement la représentation centrée objet.

(i) En systématique: identification d'un spécimen.

(ii) Choix et utilisation d'un modèle de croissance bactérienne suivant l'espèce et les conditions de culture.

(iii) Définition et analyse de plans expérimentaux, au sens de la statistique Fishérienne.

(v) Gestion des ressources forestières: choix d'un modèle de prévision (Inventaire Forestier National), intégration de connaissance sylvicoles.

(vi) Analyse des données: choix d'une méthode, aide à l'exploitation des résultats dans un domaine d'application, par exemple en Ecologie.

(vi) Guidage en régression non-linéaire, exploitation des dosages biologiques, par exemple, pour la méthode ELISA.

(vii) Pilotage ou gestion d'un système faisant intervenir modèles mathématiques, souvent utilisés quantitativement, et connaissances qualitatives, (par exemple, pilotage d'un procédé "biotechnologique", gestion de ressources naturelles), et/ou des objets vus sous différentes optiques (par exemple un objet biologique peut être vu sous l'angle phylogénétique, ou comme élément d'un système écologique, agronomique ou biotechnologique).

On peut regrouper ces questions en trois grands types:

- celles relevant du point (i), il s'agit d'identifier un objet (spécimen) dans une classification (de type plus ou moins phylogénétique), c'est-à-dire essayer d'attacher une instance décrite à un haut niveau dans une hiérarchie de schémas à un schéma de classe situé le plus bas possible dans cette hiérarchie (nous proposons d'appeler ce type de raisonnement "identification dans une classification" ou plus simplement "classement");

- celles relevant d'une activité plus spécifiquement Edora tournant autour du choix d'un modèle et/ou d'une méthode, dans un champ de connaissance méthodologique ou biologique particulier; on peut d'ailleurs remarquer qu'au moins dans certains cas, un raisonnement de type identification pourrait être aussi appliqué;

- celles où il est nécessaire de manipuler des connaissances qui se distinguent au niveau du formalisme (modèles mathématiques, et connaissances non "mathématisées"), à celui de l'implantation (déclarative ou procédurale) ou par leur la nature (mélange de connaissance de diverses disciplines: mathématiques, biologiques...) ou enfin par la façon de voir un objet (un même objet peut être examiné sous différents points de vue).

Enfin, il sera nécessaire d'envisager l'implantation de raisonnements pouvant traiter des problèmes "spécifiques", par exemple le raisonnement spatial. Ceux-ci sont le plus souvent suggérés par le domaine d'utilisation, mais ils sont formalisés et implantés par les informaticiens, ce qui met encore une fois en évidence l'importance de l'approche pluridisciplinaire. La classification intégrée dans SHIRKA est un bon exemple de cette démarche dans la mesure ou elle a été initialisée par des problèmes posés par la systématique. Cette approche (pluridisciplinaire) demande malgré tout à la fois une ouverture d'esprit des partenaires et une bonne culture de chacun d'eux dans la discipline de l'autre.

PUBLICATIONS ET COMMUNICATIONS

Publications dans des revues ou ouvrages collectifs

[1] **Rechenmann F.** - La construction de modèles dynamiques. In *"Analyse de Système en Géographie"*, Ed. Germond Y., Presses Univ. de Lyon, 1984.

[2] **Gouzé J.L.** et **Vignard P.** - EDORA: un système intelligent d'aide à la modélisation. *AMSE Review*, 2, 3, 9-26, 1984.

[3] **Corman A., Carret G., Pavé A., Flandrois J.P., Couix C.** - Bacterial growth measurement using an automated system: mathematical modelling and analysis of growth kinetics. *Ann. Inst. Pasteur / Microbiol.*, 1986, 137-B, 133-143.

[4] **Huet S. et Messéan A.** - A generalization of Gauss-Newton and Gauss-Marquardt algorithms for general estimation problems. *Computational Statistics Quarterly*, 1986.

[5] **Pavé A. et Rechenmann F.**- Computer aided modelling in biology: an Artificial Intelligence approach. In *"A.I. Applied to Simulation"* , Ed. Kerckhoffs, Vansteenkiste G.C., Zeigler B.P.. *SCS Simul. Serie*, 18, 1986, 52-66.

[6] **Pavé A., Corman A., Bobillier-Monot B.**- Utilisation et interprétation du modèle de Gompertz, application à l'étude de la croissance de jeunes rats musqués (Ondatra zibethica L.). *Biom. Praxim.*, 1986, 26, 123-140.

[7] **Rechenmann F.** - Représentation des connaissances dans les logiciels de calcul scientifiques. In *"Informatique et Calcul, Computers and Computing"*. Ed. Chenin P., Di Crescenzo C., Robert F., Masson & Wiley, 1986.

[8] **Pavé A.** - Systèmes experts: application à la modélisation en biologie. *MATAPLI*, 11, 5-19, 1987.

Thèse

Rousseau B. - Vers un environnement de résolution de problèmes en biométrie. Apport des techniques de l'intelligence artificielle et de l'interaction graphique. *Thèse de l'Université de Lyon*, formation doctorale "Analyse et modélisation des systèmes biologiques", 1988.

Travaux publiés dans cette première issue des Cahiers d'Edora

[9] **Cheruy A.** - Méthodologie de la modélisation.

[10] **Pavé A.** - Interprétation et construction de modèles différentiels de la dynamique des populations à l'aide de schémas fonctionnels.

[11] **Houllier F.** - Construction et interprétation de modèles dynamiques: Exemples forestiers.

[12] **Gentil S.** - Modélisation et indentification en Automatique - Transferts possibles vesr les bio-systèmes.

[13] **Messéan A.** - Identification de modèles dynamiques: aspects statistiques.

[14] **Houllier F.** et **Pavé A.** - Estimation initiale des paramètres d'un système différentiel linéaire en fonction des paramètres. Application aux modèles de courbes de croissance.

[15] **Rousseau B.** et **Rechenmann F.** - Le projet Edora: Vers un poste de travail informatique pour l'aide à la modélisation des systèmes dynamiques en biologie.

[16] **Gouze J.L.** et **Sciandra A.** - La loi exponentielle et ses vérifications expérimentales en biologie.

[17] **Baratchart L., Benoit E., Grimm J.** - Redressabilité des champs quadratiques plans sans singularité.

[18] **Pierret-Golbreich C.** - Vers l'interprétation des objets symboliques et biologiques dans Edora.

Colloques (avec Actes et Sélection)

[19] **Bensaïd A., Granier D.** - Rechenmann F. - SHIRKA: des systèmes experts centrés-objet". *Les Systèmes Experts et leurs Applications*. ADI, Avignon, 1984.

[20] **Rechenmann F.** - Intelligence Artificielle et construction de modèles dynamiques". *Intelligence Artificielle et Productique*, Paris, 1984.

[21] **Gentil S. et Rechenmann F.** - Identification des procédés et intelligence artificielle. *Congrès AFCET Automatique*, Toulouse, 1985.

[22] **Rechenmann F.** - SHIRKA: mécanismes d'inférence sur une base de connaissances centrée-objet". 5ème Congrès-Exposition AFCET-ADI-INRIA *"Reconnaissance des formes et Intelligence artificielle"*. Grenoble, 1985.

[23] Bensaïd A, Rechenmann F., Simonet A., Vignard P. - Mécanismes d'inférences et d'explication dans les bases de connaissances centrées-objet. *Bases de Données et bases de Connaissances*, 8ème Journées Francophones sur l'Informatique, Grenoble, 1985.

[24] **Huet S. et Messéan A.** - NL: a statistical package for general nonlinear regression problems. *Proceed. of the Compstat 86 Conf.*, Ed.De Antoni F, Lauro N., Rizzi A., 1986, 326-331.

[25] **Rousseau B., Pavé A., Rechenmann F. et Landau M.** - Edora project: Artificial Intelligence approach and work station concept to aid dynamic modelling in biology and ecology. *Suppl. Proceed. of the Summer Computer Conference*, Reno Nevada, 1986, SCS, 14-20.

[26] **Pavé A.** - Schémas fonctionnels et modélisation. Etude de modèles de la dynamique des populations. Actes du Coll. *"Biométrie-Econométrie"* , Ed. Demongeot J. et Malgrange P., sous presse (presses Univ. Dijon), 1986.

[27] **Rechenmann F., Doize M.S.** - SAFIR-SHIRKA: un système à base de connaissances centrées objet pour l'analyse financière. 7ème journées Internationales *"Les Systèmes Experts et leurs Applications"*, Avignon, 1987.

[28] **Aguirre Cervantès J.L., Bloch D., Rechenmann F. et Rouibah N.** - SHIRKA: compilateur, explication et cohérence dans des bases de connaissances centrées-objet. *MARI 87*, Paris, 1987.

[29] **Comby S., Flandrois J.P., Pavé A.** - A contribution to the study of discriminant capacity of expert system. Application to the clinical bacteriology. *Proceed. of the conf. A.I. and Cognitive Sciences*, Grenoble, 1987 (sera publié par Manchester Univ. Press).

[30] **Pierret-Golbreich C.** - Centered Knowledge Representation for Modelling in Biology. *Proceed. Internat. Symp. on AI, Expert Systems and Language in Modelling and Simulation*, Barcelone, 1987.

Rapports - Notice

[31] **Rechenmann F., Vignard P.** - CRIKA: quand les règles rencontrent les schémas. Rapport de Recherche INRIA, Sophia-Antipolis, 1985.

[32] **Pavé A., Lebreton J.D.** - Biologie de Populations: outils informatiques d'aide à la modélisation. Rapport A.R.U., 1985.

Communications

Gouzé J.L., Rechenmann F. et Vignard P. - EDORA: An Artificial Intelligence Approach to Dynamic System Modelling. *The Management an Modelling of Dynamic Systems*, Bruges,1984.

Gouzé J.L. et Vignard P. - Intelligence Artificielle et Modélisation en Biologie - Coll. COGNITIVA, Paris, 1984.

Gouzé J.L. et Vignard P. - EDORA: un système intelligent d'aide à la modélisation. *International 84 AMSE Conf. "Modelling and Simulation"*, Athen 1984.

Pavé A. - Schematic représentation: an aid to mathematical modelling and model interpretation in biology. *Intern. Working Conf. on Artificial Intelligence in Simulation*, Gand, 1985.

Pavé A. et Rechenmann F. - Object Oriented Knowledge Representation. *Intern. Working Conf. on Artificial Intelligence in Simulation*, Gand, 1985.

Rechenmann F. - Représentation des connaissances centrée-objet et modélisation dynamique: Edora. Séminaire INRIA-CNRS *"Intelligence Artificielle et Génie Moléculaire"*, INRIA, Sophia-Antipolis, 1986.

Rousseau F., Rechenmann F. - Le projet Edora: vers un poste de travail informatique pour l'aide à la modélisation des systèmes dynamiques en Biologie. 13ème Coll. International d'Econométrie Appliquée *"Aux Frontières de l'Econométrie: Expériences en Biologie et Econométrie"*, Sophia-Antipolis, 1986.

Conférences invitées

Pavé A. - Systèmes experts et modélisation. Coll. APRI et MEDIMAT, MRT , Paris, 1986.

Pavé A. - Modèles Mathématiques de la Dynamique des Populations: étude de leurs relations pour l'organisation d'une base de connaissances. Université de Pau, Séminaires de l'IBEAS, 1986.

Pavé A. - Outils Informatiques d'Aide à la Modélisation. Université de Pau, Séminaires de l'IBEAS, 1987.

Rechenmann F. - Bases de Données - Bases de Connaissances, Application à Shirka. Ecole Polytechnique Fédérale de Lausanne, 1987.

Rechenmann F. - Représentation de la connaissance dans Edora. Journées Systèmes Experts et Biométrie, INRA, INA-PG, Paris 1987.

Messéan A. - Edora: un système expert d'aide à la modélisation. Journées Systèmes Experts et Biométrie, INRA, INA-PG, Paris 1987.

Rechenmann F. - Développements récents dans SHIRKA. Séminaire CRISS, Grenoble, 1987

Pavé A. - Intelligence Artificielle et Calcul Scientifique. INA-PG, 1987.

Pavé A. - Knowledge Based Systems: Applications in Biology and Ecology (Mathematical Modelling and some other topics). University of Cambridge, U.K., (Conf. of the Biological Information Processing Group).

Liste des participants au projet et au Club Edora

1. Projet

INRIA
Centre de Sophia Antipolis
06560 VALBONNE

MM L. Baratchart, P. Bernhard, J.L. Gouzé, J. Grimm, C. Lobry (et Université de Nice)
INRIA Centre de Rocquencourt
Domaine de Voluceau, B.P. 105
78153 LE CHESNAY Cedex

Mme C. Pierret-Golbreich
INRIA / Laboratoire ARTEMIS
USMG, B.P. 68
38402 St Martin d'Hères

M F. Rechenmann
INRA
Centre de Recherche de Jouy-en-Josas
78350 JOUY-EN-JOSAS

Mlle A. Bouvier, Mlle F. Gielis, M. A. Messéan
Laboratoire de Biométrie
Université Claude Bernard (Lyon 1)
69622 VILLEURBANNE Cedex

Mme **N. Gautier**, MM. **J.C. Hervé, F. Houllier** (en poste à l'Inventaire forestier National, Montpellier), **A. Pavé, B. Rousseau** (en détachement au laboratoire ARTEMIS à Grenoble)

2. Club

Président: **A. Pavé**

Ce Club regroupe les acteurs du projet plus les collègues suivants:

M.	**C. Bernstein** (Biométrie Lyon)
Mme	**A. Chéruy** (Laboratoire d'Automatique de Grenoble, UA CNRS)
MM	**J. Demongeot** (Laboratoire de Statistique et d'Informatique Médicales, USM Grenoble)
	E. Fiolitakis (Institut für Biotechnologie, Jülich RFA)
	P.H. Gouyon (CEPE-CNRS, Montpellier)
	J. Henry (INRIA Centre de Rocquencourt)
	H. Johannes (CNRF - Biométrie - Seichamps)
	E. Jolivet (Biométrie INRA CRJJ, Jouy-en-Josas)
	J. Langla (ENSAM, Talence)
	J.D. Lebreton (CEPE-CNRS, Monpellier)
Mme	**F. Mazat** (ENSAM, Talence)
MM	**J.P. Mazat** (IBCN-CNRS, Bordeaux)
	Nival (Station Marine, Villefranche sur mer)
Mme	**E. Pommies** (INA-PG, Paris)
Mlle	**C. Reder** (Université Bordeaux, Talence)
MM	**C. Ripoll** (Université de Rouen, Lab. Echanges Cellulaires. UA CNRS, Mont St Aignan)
	S. Strizyk (Ing. Conseil, Paris)
	M. Thellier (Université de Rouen, Lab. Echanges Cellulaires. UA CNRS, Mont St Aignan)
	J. Wolsack (ENGREF, Paris)

Ce Club se réunit une fois par an à l'INRIA (Centre de Sophia-Antipolis), généralement début juillet, cette réunion se présente sous la forme d'un forum. Toute personne intéressée par l'activité de ce club peut prendre contact avec son président.

MÉTHODOLOGIE DE LA MODÉLISATION

Arlette CHERUY
Laboratoire d'Automatique
UA CNRS 228
ENSIEG - BP 46
38402 Saint Martin d'Hères

La modélisation est une technique largement utilisée dans beaucoup de disciplines: de plus, en plus, les modèles mathématiques sont reconnus comme des outils intéressants voire indispensables pour l'analyse, la commande ou encore l'aide à l'expérimentation et à la conception de systèmes ou procédés. Cependant, en biologie, la conception et l'utilisation de modèles restent problématiques pour plusieurs raisons:

1) Les systèmes à modéliser sont mal maîtrisés et on peut même se demander l'intérêt de représenter mathématiquement un système mal connu si on ne souligne pas l'utilité d'un modèle pour tester des hypothèses ou pour mieux appréhender un comportement.

2) Les systèmes biologiques sont en général difficiles à "manipuler", à expérimenter; les mesures sont souvent délicates et/ou complexes et les capteurs font défaut.

3) L'élaboration d'un modèle fait appel à des compétences variées: biologique, mathématique, informatique voire automatique qui sont rarement réunies au sein d'une même équipe ou laboratoire en particulier de biologie. Aussi, jusqu'à maintenant, les études de modélisation se font "au coup par coup" et prennent des aspects différents suivant les compétences que l'on a pu réunir.
En effet, pour élaborer un modèle, il est tout aussi nécessaire de comprendre le phénomène biologique à modéliser, de connaître l'expérimentation, ses possibilités et limites, que d'avoir des idées sur les différentes formulations mathématiques possibles et leurs propriétés, que de maîtriser les outils numériques permettant de mettre en œuvre la simulation ou l'identification d'un modèle.

Pour avoir été impliqués dans de nombreuses études de modélisation de systèmes biologiques, (cinétiques enzymatiques, fermentations...), nous avons acquis, par la pratique, des connaissances dans les différents domaines qui nous ont amenés à une certaine réflexion sur l'approche modélisation et son impact en biologie si bien qu'il nous apparaît maintenant possible de proposer une méthodologie de la conception d'un modèle et de son "bon usage" en fonction des problèmes posés et des informations disponibles (données expérimentales, connaissances a priori...).

Cette méthode peut être présentée sous forme d'un organigramme (figure 1) où l'on définit les différentes étapes dans l'élaboration d'un modèle et leurs imbrications. Ces étapes sont au nombre de 7, ce sont:

1) La définition des objectifs

Il est très important de définir au départ l'objectif de la modélisation car un modèle ne peut être conçu que pour un objectif précis et en fonction de son utilisation ultérieure. En effet, la conception d'un modèle général répondant à des objectifs multiples n'est pas réaliste, et d'autre part, il est hasardeux d'utiliser un modèle dans une

application pour laquelle il n'a pas été conçu. De plus, la prise en compte de l'objectif peut être déterminante pour la suite, par exemple au niveau de la formulation mathématique: un objectif de commande dynamique de procédé implique pratiquement de retenir un modèle déterministe et linéaire car c'est le seul type de modèle pour lequel on sache calculer théoriquement une commande.

2) L'analyse du système

C'est une étape essentiellement "qualitative" où il s'agit de délimiter le système et son environnement, de définir les variables internes significatives (variables d'état du système) ainsi que leurs interactions. Il convient donc de répertorier la connaissance a priori sur le système et les possibilités d'expérimentation. Le résultat de cette étape d'analyse doit pouvoir se concrétiser par un schéma fonctionnel du système étudié, visualisant l'ensemble des variables et de leurs interactions.

Cette analyse du système s'appuie sur des données expérimentales et la connaissance des spécialistes et l'on est confronté à ce niveau souvent à un problème de tri, de structuration, de synthèse de l'information.

3) Formulation mathématique

C'est la phase de mise en équation, il s'agit simplement de traduire mathématiquement les interactions retenues au niveau du schéma fonctionnel.

Ainsi, on est amené à faire des choix techniques, par exemple, entre modèle déterministe ou stochastique, entre modèle linéaire ou non linéaire, entre modèle continu ou discret...

Pour faire ces choix, une bonne connaissance des propriétés des différents modèles et de leurs possibilités d'application est nécessaire.

4) Test de modèle

Il s'agit à ce niveau de tester les propriétés du modèle et de vérifier si son comportement est cohérent avec les expérimentations dont on dispose avant de passer à l'identification des paramètres.

C'est à ce niveau qu'il convient de faire une analyse de sensibilité, qui est très importante, en particulier pour détecter les paramètres difficiles à identifier (c'est-à-dire ceux qui influent peu sur les grandeurs mesurées) et pour guider l'expérimentation, par exemple, en définissant les meilleurs instants de mesure (c'est-à-dire ceux pour lesquels la sensibilité est maximum).

5) Expérimentation

Elle doit être simultanée à la construction du modèle car il est peu vraisemblable que l'on puisse établir un modèle sur un jeu de données obtenues préalablement à toute réflexion sur les objectifs de la modélisation et sur la structure envisagée pour le modèle. Et de plus, l'expérimentation nécessaire peut évoluer avec l'élaboration du modèle.

6) Identification

Il s'agit ici d'estimer les paramètres du modèle retenu à partir des données expérimentales.

7) Validation du modèle

Un modèle doit être validé avant d'être utilisé. Un bon ajustement entre données expérimentales et simulées n'est pas suffisant pour considérer le modèle comme prêt pour l'application.

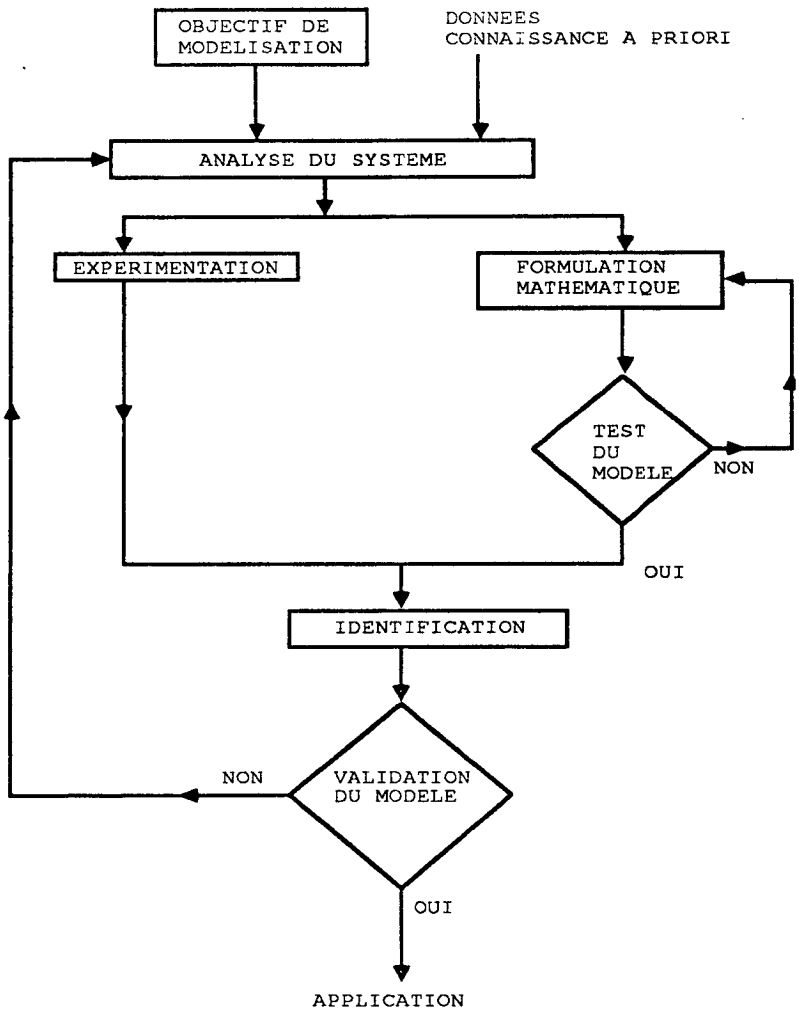


Figure 1 - Méthodologie de la modélisation

I - Les objectifs de la modélisation

L'objectif de la modélisation doit être clairement défini au départ. Un modèle général qui simulerait toutes les propriétés d'un système n'est pas réaliste. Et d'autre part, il est illusoire de vouloir utiliser un modèle pour une application différente de celle pour laquelle il a été conçu.

Les principaux objectifs de la modélisation des systèmes biologiques sont:
- le test d'hypothèses ou de mécanismes,
- la simulation, la prédiction, la prévision,
- l'estimation de paramètres non directement mesurables,
- la commande des procédés, l'optimisation, la commande optimale...
- l'aide à l'expérimentation: optimisation d'expériences...

1) Test d'hypothèse ou de mécanisme

Dans ce cas, un modèle est construit pour tester si une hypothèse sur la structure ou le fonctionnement d'un système biologique est compatible avec les observations et données expérimentales. Par exemple, dans les systèmes biologiques qui mettent en œuvre de nombreuses réactions enzymatiques, il est intéressant de tester si une hypothèse d'inhibition ou d'activation est compatible avec les résultats cinétiques observés.

La démarche est la suivante:
a) A partir des connaissances a priori sur le système à étudier, on formule des hypothèses sur son fonctionnement ou sa structure.
b) On construit un modèle qui prend en compte ces hypothèses et les traduit mathématiquement.
c) On utilise ce modèle pour simuler le comportement du système en particulier dans les conditions où l'on peut l'expérimenter.
d) On compare résultats expérimentaux et simulés afin d'apprécier la validité de l'hypothèse de départ.

Il est important de souligner qu'il s'agit là d'un test et non d'une démonstration, c'est-à-dire que le résultat indiquera simplement si l'hypothèse en question est (ou n'est pas) compatible avec les données expérimentales considérées. De plus, comme une hypothèse biologique peut avoir plusieurs formulations mathématiques, le test concerne en réalité une certaine formulation mathématique d'une hypothèse biologique.

2) Simulation

L'intérêt de la simulation en analyse de systèmes n'est plus à démontrer en particulier pour obtenir l'évolution simultanée de toutes les variables internes d'un système même celles qui ne sont pas directement mesurables.

Cela nécessite simplement que le modèle ait comme variables d'état, les variables internes du système. Cette simulation n'est valable que si les hypothèses qui ont servi à l'élaboration du modèle sont vérifiées, en particulier, elle ne concerne que le domaine de temps correspondant aux mesures et observations utilisées pour la modélisation. Si on utilise le modèle pour simuler le fonctionnement du système pour des temps postérieurs à ceux de l'observation ou des mesures, on fait alors de la prédiction. Et si on considère des conditions expérimentales différentes (présence de perturbations par exemple), on fait de la prévision. Prédiction et prévision ne sont valables que si les hypothèses sur le système sont les mêmes que durant les observations qui ont servi à élaborer le modèle.

3) Estimation de paramètres non directement mesurables

Cet objectif est très intéressant en biologie car les expériences sont souvent difficiles et les capteurs limités. Dans ce cas, le modèle doit avoir comme paramètres, les paramètres biologiques à estimer. Exemple: estimation du taux de croissance de bactéries dans un procédé biologique.

4) Commande de procédé, optimisation, commande optimale

Un modèle est nécessaire pour définir les conditions optimales de fonctionnement d'un procédé et pour déterminer les commandes qui le maintiennent en permanence dans cet état optimal de fonctionnement.

Dans ce cas, le modèle doit prendre en compte essentiellement les relations de cause à effet entre variables de sortie (représentant par exemple une production à maximiser) et variables de commande (qui sont des variables que l'on peut maîtriser comme un débit d'alimentation, un pH, une température...). Les variables d'état du modèle ne représentent pas nécessairement des variables internes du système biologique.

5) Aide à l'expérimentation – optimisation d'expérience

Comme les expériences sont souvent difficiles à mettre en œuvre, il est important de pouvoir les planifier de manière à en tirer le maximum d'informations. Un modèle est nécessaire pour concevoir un plan d'expérience. Par exemple, si le nombre de prises d'échantillons est limité parce que les mesures nécessitent une analyse biologique complexe, il est intéressant de pouvoir déterminer quels sont les meilleurs instants d'échantillonnage (c'est-à-dire ceux qui permettent d'avoir un maximum d'informations).

II - L'analyse du système

Le but de cette étape est de répertorier les connaissances disponibles et les expériences possibles sur le système biologique à analyser de manière à dégager un schéma fonctionnel visualisant toutes les interactions entre constituants de base à prendre en compte dans le modèle

Dans cette étape, on devra donc:

- délimiter le système de son environnement;
- dégager les éléments de base du système, les variables significatives;
- définir leurs interactions;

et ce, de manière à pouvoir exprimer le résultat sous forme de schéma fonctionnel.

Exemple de schéma fonctionnel:

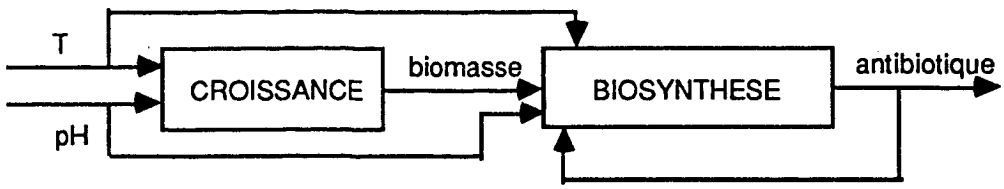


Schéma fonctionnel d'un procédé de fermentation
où l'on s'intéressait essentiellement à une production maximale d'antibiotique
en agissant sur la température et le pH
et où l'on ne souhaitait pas considérer le métabolisme intermédiaire.

1) Analyse des données

Cette analyse est une étape délicate qui s'appuie sur les données expérimentales et les connaissances des spécialistes qui sont souvent dispersées et touffues. Aussi, on est confronté à un problème de tri, de structuration, de synthèse de l'information. Si l'information et les données dont on dispose sont importantes, on peut faire appel aux techniques d'analyse de données, par exemple, pour dégager les variables significatives. Ces méthodes d'analyse de données sont issues de la statistique, elles recherchent et analysent systématiquement les corrélations et les relations entre variables. Il en existe de trois types.

* *des méthodes descriptives* (ex: analyse en composantes principales, analyse de correspondance) qui permettent d'avoir une représentation résumée, plus synthétique de l'information.
* *des méthodes explicatives* (ex: analyse canonique, analyse discriminante) qui permettent d'étudier la significativité des variables, de savoir si l'on peut expliquer les variations d'une variable par celles de telle ou telle autre.
* *des méthodes de classification de structuration des données* qui permettent de classer des données, de former des groupes.

Il existe des logiciels spécialisés qui rendent leur mise en œuvre aisée. Un exemple d'application est donné en annexe I où l'analyse des correspondances a été utilisée pour dégager des variables significatives du système cholestérol.

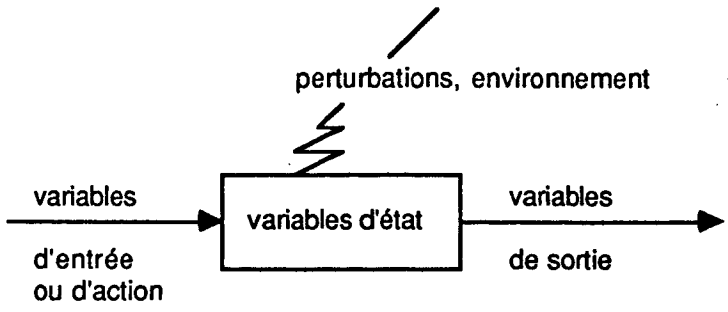
2) Délimitation du système, définition de l'environnement

Il est souvent difficile de délimiter un système biologique de son environnement et de caractériser cet environnement. Les systèmes physiologiques par exemple, sont très interconnectés donc difficiles à isoler. On ne sait pas où couper les chaînes métaboliques.

Classiquement sont incluses dans l'environnement des variables comme la température, le pH, des perturbations, des grandeurs qui peuvent agir sur le système mais qui ne subissent pas son influence.

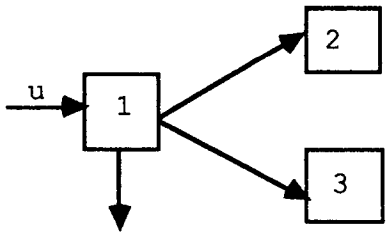
3) Définition des variables

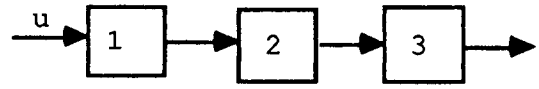
On peut définir trois types de variables:



* *Les variables d'entrée*, extérieures au système, sur lesquelles on peut agir pour contrôler son fonctionnement. Le choix de ces variables est très important quand l'objectif de la modélisation est la commande de procédé; il convient alors de prendre en compte les propriétés de commandabilité du système. (Un système est dit commandable par une entrée quand il est possible de trouver les valeurs de cette entrée qui permettent de faire passer le système d'un état à un autre en un temps fini. Pour les systèmes modélisés par un système d'équations différentielles linéaires, il existe une condition nécessaire et suffissante de commadabilité).

Exemple: commandabilité de systèmes à compartiments:





$$\frac{dx_1}{dt} = -k_{12}x_1 - k_{13}x_1 - k'_{11}x_1 + k_{01}u$$

$$\frac{dx_2}{dt} = k_{12}x_1$$

$$\frac{dx_3}{dt} = k_{13}x_1$$

$$\frac{dx_1}{dt} = -k_{12}x_1 + k_{01}u$$

$$\frac{dx_2}{dt} = k_{12}x_1 - k_{23}x_2$$

$$\frac{dx_3}{dt} = k_{23}x_2 - k_{30}x_3$$

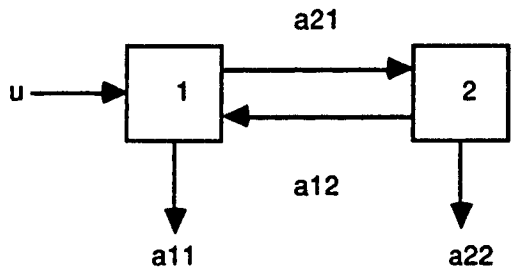
u variable d'entrée
k_{ij} = constantes
x_i variable d'état

système non commandable
quelles que soient les valeurs des paramètres k_{ij}, x_2 sera toujours proportionnel à x_3.

système commandable

* *Les variables de sortie*, elles permettent de mesurer l'état du système. Souvent les variables de sortie sont directement les variables d'état du système, mais elles peuvent en être une combinaison ou une fonction plus ou moins complexe. Si un choix est possible, il doit être guidé par les propriétés d'observabilité du système. Un système est observable par une variable de sortie, si, celle-ci permet d'observer n'importe quel état du système.

Exemple:



variables d'état x_1 et x_2 ; a_{ij} = constantes

$$\frac{dx_1}{dt} = au - (a_{11} + a_{21})x_1 + a_{12}x_2$$

$$\frac{dx_2}{dt} = a_{21}x_1 - (a_{12} + a_{22})x_2$$

Si on considère comme sortie, les variables d'état x_1 et x_2, le système est observable $\forall$ aij. Par contre, il sera non observable dans les trois cas suivants:

- si la sortie est x_1 et si $a_{12} = 0$
- si la sortie est x_2 et si $a_{21} = 0$
- si la sortie est $(x_1 + x_2)$ et si $(a_{11} + a_{21}) = 0$ ou $(a_{12} + a_{22}) = 0$

Cette propriété d'observabilité est fondamentale à considérer pour la conception de l'expérimentation, pour le choix des variables à mesurer.

* *Les variables d'état*: ce sont les variables internes du système dont la connaissance définit complétement l'état du système. Généralement pour les systèmes biologiques, les variables d'état représentent des éléments comme les concentrations de certains réactants. La définition de ces variables nécessite d'avoir choisi le niveau d'agrégation à considérer (niveau moléculaire, cellulaire, population...); choix qui résulte d'un compromis entre une simplicité irréaliste et une complexité non vérifiable.

Exemple: dans un procédé de fermentation, où des microorganismes sont utilisés pour produire un antibiotique, il est évident que l'activité des microorganismes est une variable d'état mais tous les microorganismes ne sont pas dans le même état physiologique, certains sont jeunes et très actifs, d'autres sont vieux et peu actifs. Pour tenir compte de cet état, certains auteurs ont proposé des modèles où la biomasse est structurée en âges. Mais on aboutit à des modèles complexes pratiquement inutilisables à des fins de commande. Aussi, on utilise généralement une seule variable d'état représentant la concentration en microorganismes.

Cependant, une seule variable d'état peut ne pas être suffisante même si l'on ne veut pas tenir compte des différents états physiologiques. Dans une de nos études, nous avons dû considérer deux variables d'état pour pouvoir rendre compte de certaines propriétés de mémoire de microorganismes.

Quand l'objectif de la modélisation est la commande de procédé, on s'intéresse essentiellement à des relations de cause à effet entre variables d'entrée et de sortie et dans ce cas, les variables d'état n'ont pas à représenter obligatoirement des variables biologiques, ce ne sont que des outils intermédiaires pour caractériser les relations entrée-sortie. Et quand, de plus, les modèles sont linéaires, il existe des techniques pour déterminer le nombre minimal de variables d'état nécessaires (problème de réalisation minimale).

Dans cette phase de définition des variables, les possibilités d'expérimentation doivent être répertoriées car la compléxité d'un modèle, son degré de "finesse" en dépendent étroitement.

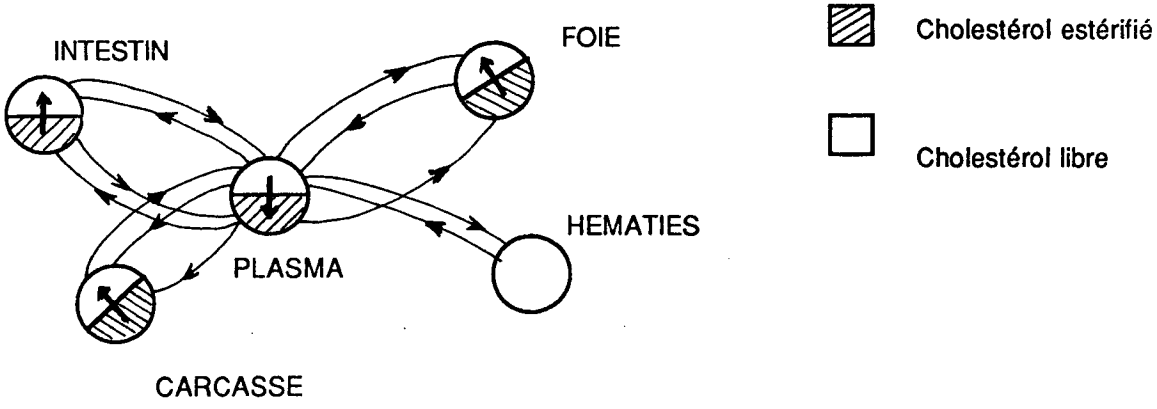
4) Définition des relations entre variables, schéma fonctionnel

Finalement, pour pouvoir établir un schéma fonctionnel, il reste à définir les relations entre variables à considérer.

Par exemple, lorsque l'on étudie un système métabolique, ou de transport, il est souvent intéressant d'utiliser l'analyse compartimentale. Cette approche est couramment employée en biologie, en particulier, en pharmacocinétique ou quand des traceurs sont utilisés pour l'expérimentation.

Un compartiment est défini, comme un réservoir unitaire où peut s'accumuler un réactif et chaque compartiment est caractérisé par sa concentration en réactif. Ainsi définir des compartiments et des flux entre ces compartiments revient à établir un schéma fonctionnel.

Exemple: le système cholestérol: approche compartimentale:

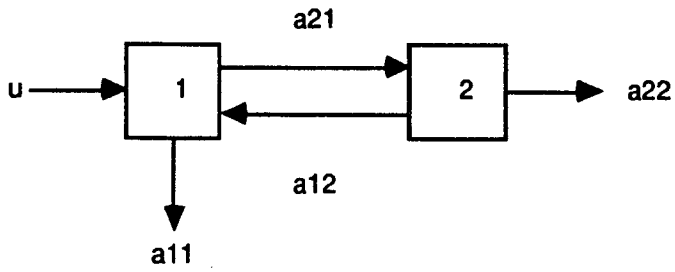


- Système à cinq compartiments avec en général deux variables d'état par compartiment représentant les deux formes sous lesquelles peut se présenter le cholestérol.
- Les compartiments représentant en général des organes.

III - Elaboration du modèle, formulation mathématique

Dans cette étape, il s'agit d'écrire les équations du modèle, c'est-à-dire de traduire mathématiquement les relations définies lors de l'analyse du système. Le problème sera très différent suivant que l'on s'intéresse aux relations de cause à effet entre entrées et sorties (modèle de représentation) ou aux relations entre variables d'état du modèle représentant des variables physiques ou biologiques (modèle de connaissance).

En biologie, le second cas est plus fréquent; dans ces conditions, les équations expriment généralement des bilans de matière ou d'énergie.
Par exemple, dans un système à compartiment, le modèle est obtenu en écrivant une équation de bilan pour chaque compartiment: quantité de réactif emmagasinée par unité de temps dans un compartiment = somme des flux entrants – somme des flux sortants; les flux étant exprimés en fonction des variables d'état (en général, fonction linéaire de la concentration du réactif dans le compartiment d'origine)



x_1 et x_2, variables d'état, concentrations du réactif dans les compartiments 1 et 2 respectivement.

$$\frac{dx_1}{dt} = a_{12}\,x_2 - (a_{21} + a_{11})\,x_1 + u$$

$$\frac{dx_2}{dt} = a_{21}\,x_1 - (a_{22} + a_{12})\,x$$

Dans cet exemple, les flux sont supposés proportionnels à la concentration dans le compartiment d'origine; aussi, on aboutit à un système d'équations différentielles linéaires.

La formulation mathématique requiert plusieurs choix techniques:

- *Modèle déterministe ou stochastique*

Etant données les caractéristiques des systèmes biologiques, on est tout naturellement orienté vers des modèles probabilistes; or, ces modèles sont très difficiles à utiliser aussi retient-on plutôt les modèles déterministes qui travaillent sur les valeurs moyennes.

- *Modèle linéaire ou non linéaire*

Ce choix dépend de l'application que l'on veut faire du modèle et des phénomènes dont on veut rendre compte. D'un côté, un modèle linéaire est facile à utiliser, la théorie est très développée, il est bien adapté pour un calcul de commande. D'un autre côté, les modèles non linéaires représentent souvent mieux la réalité biologique, mais il y peu de résultats

mathématiques sur leurs propriétés et leurs applications, aussi l'étude analytique est souvent impossible. Néamoins, étant données les performances des techniques numériques, ils sont d'usage courant pour un objectif de simulation.

- *Modèle continu ou discret (dans le temps)*

Ce choix doit être fait en tenant compte que:

* D'un côté, les modèles discrets sont très faciles à simuler car ils sont définis par des équations aux différences $x_{k+1} = f(x_k, u_k, p)$ et ne nécessitent donc pas d'algorithme d'intégration numérique. Cependant, le temps est échantillonné et les paramètres des équations aux différences dépendent de la période d'échantillonnage. Les modèles discrets sont donc à utiliser quand les mesures et données sont échantillonnées avec une période constante. Mais le choix de cette période est souvent problématique car aucun évènement important ne doit pas se produire entre deux instants d'échantillonnage. Quand le système est linéaire, il y a correspondance entre le modèle continu et discret, mais, pour les systèmes non linéaires, il n'existe pas de théorie pour leur discrétisation.

* D'un autre côté, les modèles continus définis par des équations différentielles ordinaires $dx = f(x, u, p)$ ont été très étudiés mais peuvent présenter des difficultés numériques au niveau de la simulation par exemple lorsque l'on a faire à des systèmes non linéaires et mal conditionnés. C'est le cas en cinétiques enzymatiques où l'on doit prendre en compte simultanément des phénomènes rapides et lents (formation rapide du complexe enzyme-substrat et dégradation lente de ce complexe pour donner le produit de réaction). Cependant, il existe maintenant des algorithmes d'intégration performants (ex: algorithme de Gear) qui permettent souvent de surmonter ces difficultés numériques.

- *Modèles à paramètres repartis ou fixés*

Souvent, on est amené à considérer des variations de l'état du système en fonction du temps, de l'espace ou de l'âge. Aussi, on peut être conduit à décrire le système par des équations aux dérivées partielles, dont le traitement numérique est généralement lourd et difficile. Il existe des méthodes d'intégration appropriées: méthodes des éléments finis, aux différences finies qui s'appuient sur un découpage du temps ou de l'espace en petits éléments à l'intérieur desquels les paramètres peuvent être considérés comme constants. L'inconvénient est que l'on aboutit à la résolution de système linéaires de grandes dimensions. Aussi, les modèles à paramètres distribués sont assez peu utilisés en biologie.

En conclusion de cette étape de formulation mathématique, on doit disposer d'un ensemble d'équations mathématiques dont on va tester les propriétés avant de passer à l'estimation des paramètres à partir des données expérimentales.

IV - Test du modèle

Le but de cette étape est d'analyser les propriétés du modèle, de tester, en particulier, si son comportement dynamique est bien celui escompté. Ces tests sont assez évidents lorsque l'on a retenu un modèle linéaire car les propriétés des systèmes linéaires sont bien connues et se traduisent généralement par des théorèmes ou conditions nécessaires et suffisantes à vérifier (ex: CNS pour la commandabilité et l'observabilité des systèmes linéaires). Par contre, pour les modèles non linéaires, les résultats théoriques sont peu nombreux et les tests sont souvent effectués par simulation, c'est-à-dire que l'on donne aux paramètres du modèle des valeurs numériques et on simule le comportement du

système pour pouvoir l'analyser et le comparer à celui observé. Il s'agit d'une comparaison "qualitative" pas d'un ajustement précis; on vérifie par exemple que le modèle permet d'obtenir la même allure de courbe que celle observée. Ainsi, en plus du test des propriétés comme commandabilité ou l'observabilité, il convient d'effectuer au moins deux types d'analyse:

1) ***Une analyse de stabilité*** qui comprend deux aspects:

a - L'étude de la stabilité dynamique qui consiste à analyser l'évolution à long terme ($t \to \infty$) des variables mesurées (sorties du modèle) et à vérifier, par exemple, qu'elles sont bornées.

b - L'étude de la stabilité structurelle qui consiste à vérifier si certaines valeurs de paramètres n'entraînent pas des modifications de comportement de l'état et des sorties du modèle.

2) ***Une analyse de sensibilité*** qui consiste à étudier l'influence des valeurs des paramètres, des conditions initiales et des variables de commande (entrées), sur l'état et les sorties du modèle. Cette analyse est très importante, en particulier, pour détecter les paramètres difficiles à identifier (c'est-à-dire ceux qui influent peu sur les grandeurs mesurées) et pour définir les meilleurs instants de mesure (c'est-à-dire ceux pour lesquels la sensibilité est maximum). Elle est un outil de base pour l'aide à l'expérimentation et à l'identification. Elle nécessite l'étude des fonctions de sensibilité Sij.

$$Sij = \frac{\partial x_i}{\partial p_j}$$

Sij représente la sensibilité de la variable d'état (ou de sortie) x_i au paramètre p_j.
Les Sij varient dans le temps et sont données par des équations du même type que celles du modèle.

Exemple:

Soit un modèle défini par n équations d'état récurrentes du type:

$$x(k+1) = f[x(k), u(k), p]$$

$\underline{x}$ vecteur d'état (dimension n)
$\underline{u}$ vecteur d'entrée
$\underline{p}$ vecteur des paramètres (dimension r)

$$Sij = \frac{\partial x_i}{\partial p_j}$$

$$Sij(k+1) = \frac{\partial f_i}{\partial p_j}(x(k), u(k), p) + \sum_{I=1}^{n} \frac{\partial f_i}{\partial x_1} slj(k)$$

qui représente un ensemble de *n x r équations* récurrentes à résoudre en même temps que celles du modèle.

NB: Les paramètres des équations de sensibilité dépendent de la valeur des paramètres du modèle.

En pratique, on estime numériquement les fonctions de sensibilité de la manière suivante:

valeur p_j du paramètre $\rightarrow$ simulation $\rightarrow$ état x_k

$p_j + \Delta p_j \rightarrow$ simulation $\rightarrow x_k + \Delta x_k$

$$Sij\,(k) = \lim_{\Delta p_j \rightarrow 0} \frac{\Delta x_i}{\Delta p_j}\,(k)$$

On peut également analyser la sensibilité en définissant des distances.

$$Dij = \left\{ \frac{1}{N} \sum_{k=0}^{N} \frac{[x_i(k) - x_{ij}(k)]^2}{x_i(k)} \right\}^{1/2}$$

x_i = variable d'état
x_{ij} = variable d'étant que la paramètre p_j a varié
N = nombre total d'échantillons (temps d'observation).

Un tel critère est intéressant car il est sans dimension et permet ainsi une comparaison. Cependant, il est également important d'analyser les fonctions de sensibilité en fonction du temps de manière à pouvoir détecter leur maximum (amplitude et position), information qui n'est pas explicitée par une distance de type Dij, mais qui est très utile pour guider l'expérimentation et l'identification.

Exemple:

Soit le modèle $\frac{dx_1}{dt} = \mu\, x_1\, (1 - \frac{x_1}{x_{1f}})$ qui décrit une croissance microbienne.

x_1 concentration microbienne
μ, x_{1f} paramètres

équations de sensibilité:

$$S_1 = \frac{\partial x_1}{\partial \mu} \quad \text{et} \quad S_2 = \frac{\partial x_1}{\partial x_{1f}}$$

$$\frac{dS_1}{dt} = x_1\,(1 - \frac{x_1}{x_{1f}}) + S_1\,(1 - 2\,\frac{x_1}{x_{1f}})$$

$$\frac{dS_2}{dt} = \frac{{x_1}^2}{{x_{1f}}^2}\mu + \mu.\ S_2\,(1 - \frac{x_1}{s_2}) - \mu\,\frac{x_1}{x_{1f}}\,S_2$$

Ensemble de deux équations différentielles à résoudre en même temps que l'équation du modèle et dont les résultats sont donnés figure 2.

On note que:

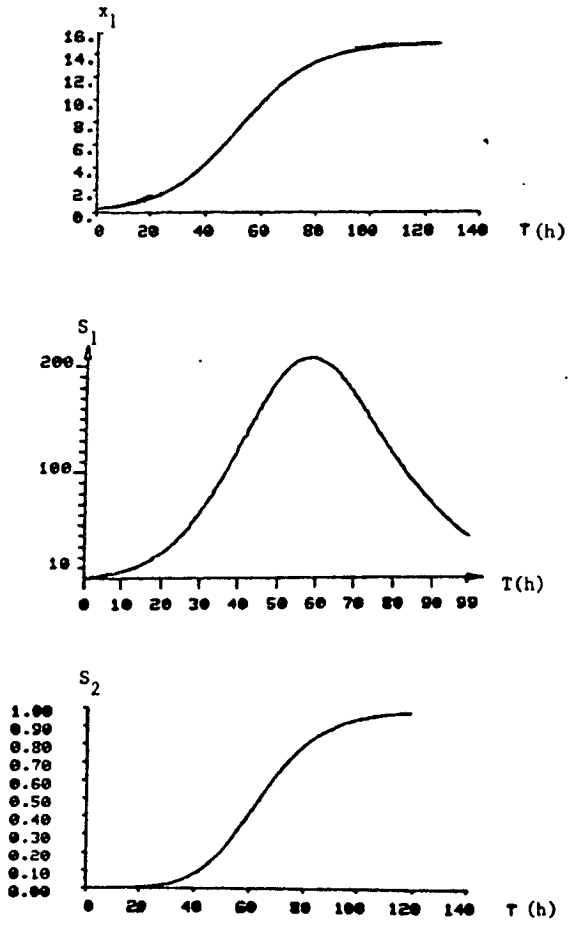


Figure 2 - Sensibilité de x_1 à μ et x_{1f}

- s_1 (t) et s_2 (t) présentent un maximum.

On aura donc intérêt à avoir de bonnes mesures de x_1 concentrées dans les régions de sensibilité maximum pour avoir une bonne identification des paramètres.

- Or ces régions de sensibilité maximum sont pratiquement confondues (50, 70h). Mais comme $S_1 >> S_2$, cette région sera surtout intéressante pour l'identification de μ.

- Lorsque $t \to \infty$, $s_1 \to 0$ et S_2 vers une valeur finie; donc $x_1(t)$ est plus sensible aux variations de x_{1f} que de μ, on pourra exploiter cette propriété pour l'identification de x_{1f}.

En conclusion, dans cet exemple, l'analyse de sensibilité nous indique que pour avoir une bonne identification de μ et x_{1f}, on aura intérêt à avoir de bonnes mesures de la croissance (x_1) dans les régions 50 h < T < 70 h et T > 90 h respectivement. Ce résultat était presque évident surtout pour le paramètre x_{1f} qui représente la valeur finale de la concentration microbienne. Cependant les modèles ne sont pas toujours aussi simples!

Pendant longtemps l'analyse de sensibilité a été peu employée car souvent l'utilisateur reculait devant l'écriture et la résolution des équations de sensibilité. Or actuellement, d'une part, il existe des méthodes de calcul formel qui permettent d'écrire

automatiquement et formellement les équations de sensibilité à partir des équations du modèle; d'autre part, on dispose de méthodes numériques performantes d'intégration pour obtenir leur solution.

Par ailleurs, il peut être également intéressant d'étudier la sensibilité des sorties du modèle aux variables d'entrée (commande) et aux variables d'état. Souvent on aborde ce problème en introduisant des composantes aléatoires.

$$u(k) = u(k) + b(k) \quad \text{ou} \quad u(k) = u(k)\,[1 + b(k)]$$

où b(k) est une composante aléatoire, blanche, centrée, gaussienne.

De même pour les variables d'état:

$$x(k+1) = f\,[x(k), u(k), p]\,[1 + b(k)]$$

En conclusion, cette étape de test du modèle reste délicate lorsque les modèles sont non linéaires, car on est amené à procéder par simulation aussi bien pour la stabilité que pour la sensibilité et l'on ne peut jamais être sûr d'avoir fait une analyse exhaustive des propriétés. Cependant ces analyses sont très intéressantes pour guider l'expérimentation et l'idendification.

V - Expérimentation

Cette étape doit être simultanée à l'élaboration du modèle car l'expérimentation doit être conçue en fonction des propriétés du modèle. Il n'est pas réaliste d'espérer identifier correctement les paramètres d'un modèle à partir d'un jeu de données obtenues préalablement à toute réflexion sur la modélisation. En biologie, il est bien connu que les mesures sont difficiles, peu nombreuses, nécessitent des analyses complexes, qu'il y a peu de capteurs et que les contraintes expérimentales sont nombreuses. Les possibilités d'excitation sont limitées: il faut se contenter d'observer les conditions normales de fonctionnement. Les systèmes biologiques sont difficiles à manipuler. Très souvent on effectue des expériences in vitro car elles sont impossibles à mettre en œuvre "in vivo"; or, l'environnement du système est différent et donc il n'y a aucune raison pour que le système se comporte de la même manière et soit caractérisé par le même modèle. Cependant les expériences in vitro sont très intéressantes pour cerner un système voir pour étudier l'influence de l'environnement, mais en aucun cas des paramètres identifiés à partir d'expériences in vitro ne peuvent être transposés sans précaution in vivo.

VI - Identification des paramètres

Il s'agit d'estimer les paramètres du modèle à partir des données expérimentales en minimisant l'écart entre résultats expérimentaux et simulés.

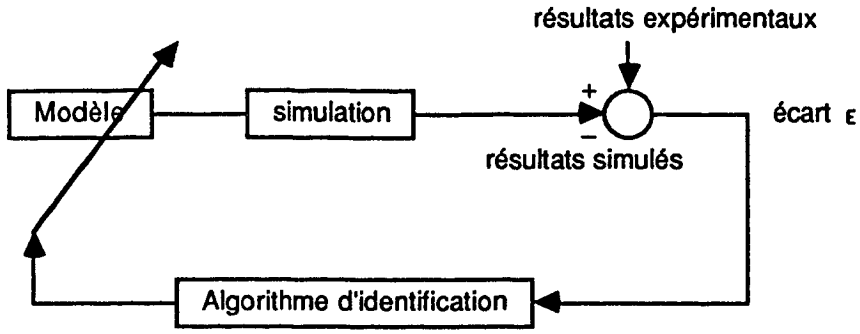


Figure 3 - Principe de l'identification

La procédure d'identification se déroule de la manière suivante.

* On donne des valeurs initiales aux paramètres du modèle.
* Par simulation, on détermine les sorties du modèle pour pouvoir les comparer aux données expérimentales.
* On détermine l'écart entre résultats expérimentaux et simulés. Soit ε cet écart.
* On définit une fonction d'erreur f (ε) que l'on cherche à minimiser en modifiant les valeurs des paramètres du modèle.

Les différentes méthodes d'identification dépendent de la fonction d'erreur et de l'algorithme d'optimisation utilisé.

Les fonctions d'erreur classiques sont:

- les moindres carrés $J = \sum \varepsilon^2$
- les moindres carrés pondérés $J = \sum \varepsilon^T W \varepsilon$

La pondération W permet, par exemple, de réduire l'inflence de certaines données peu sûres.

Maximum de vraisemblance J : P(x/p), on recherche les paramètres p qui donnent la probabilité maximum d'obtenir les valeurs observées de x.

La minimisation de la fonction d'erreur est un problème de programmation non linéaire et il existe des algorithmes disponibles dans les bibliothèques de programmes comme Harwell.

Classiquement des algorithmes type Gauss-Marquardt sont utilisés pour minimiser une fonction des moindres carrés. On a alors.

$\varepsilon = x_e - x_m$ = écart

fonction erreur = critère $J = \varepsilon^T \varepsilon$

valeur des paramètres		valeur du critère
θ_0 (valeur initiale)	---------->	J_0
$\theta_1 = \theta_0 + \Delta\theta_1$	---------->	J_1
$\theta_2 = \theta_1 = \Delta\theta_2$	---------->	J_2

Le problème est de définir les $\Delta\theta$ tels que la série $J_0 > J_1 > J_2 ....> J$ converge vers J.

Dans un algorithme de Gauss-Marquardt, les itérations se font de la manière suivante:

$$\Delta\theta_i = [S^TS + \lambda\ I]^{-1}\ S^T\ \varepsilon_i$$

où S est la matrice des fonctions de sensibilité

λ = constante

I = matrice identité

Si λ est très grand devant les valeurs propres de S^TS, on a alors un algorithme du gradient. Si λ est négligeable, on a un algorithme de Gauss Newton. Aussi, en général on fait évoluer la valeur de la constante λ au cours de l'identification = on la prend grande au départ et on la diminue à chaque itération. La procédure interactive s'arrête lorsque $J < J$ min fixé et les valeurs correspondantes des paramètres sont retenues; on calcule alors différents éléments (hessien...) pour apprécier la précision des estimations.

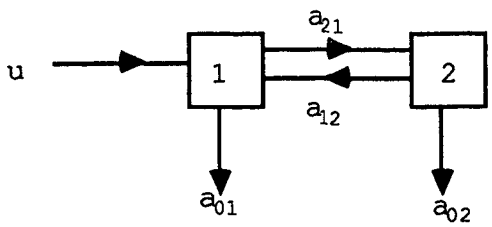
Au cours de l'identification, on peut rencontrer divers problèmes aussi bien mathématiques que numériques, en particulier, ceux liés aux propriétés d'identifiabilité des systèmes qui peuvent se situer à deux niveaux:

1) *l'identifiabilité théorique* (ou a priori) qui est une propriété du modèle mathématique, de sa structure, qui indique s'il est possible de déterminer l'ensemble des paramètres du modèle à partir des mesures des entrées u et des sorties y supposées parfaites.
   S'il existe un seul jeu de paramètres = le modèle est identifiable.
   S'il en existe plusieurs, il est localement identifiable, et s'il en existe une infinité le modèle est non identifiable.
   Si un modèle est identifiable théoriquement, cela signifie qu'il existe une solution mais ne donne pas cette solution.

2) *l'identifiabilité pratique* qui est une propriété de l'ensemble {modèle - mesures - méthodes d'identification}. Elle indique s'il est possible de trouver la valeur du jeu de paramétres compte tenu du modèle (identifiable a priori), des mesures (avec leur qualité) et des méthodes d'identification. Les causes de non-identifiabilité pratique peuvent donc venir:
   - du modèle,
   - des mesures qui sont trop imprécises, en trop petit nombre, ou faites à des instants inadéquats pour lesquels elles sont peu sensibles aux variations de paramètres...,
   - des entrées qui n'excitent pas assez le système,
   - des méthodes d'identification qui sont peu performantes ou mal adaptées...

Pour le modèle linéaire, il existe divers moyens d'étudier leur identifiabilité théorique. Par exemple, pour déterminer le nombre de paramètres identifiables d'un système, il suffit d'exprimer sa fonction de transfert; le nombre de paramètres de cette fonction de transfert indique le nombre de paramètres identifiables.

Exemple - identifiabilité d'un modèle à compartiment



entrée u

sortie $s = x_1$

$$\frac{dx_1}{dt} = -(a_{01} + a_{21})x_1 + a_{12}\,x_2 + u$$

$$\frac{dx_2}{dt} = a_{21}\,x_1 - (a_{02} + a_{12})\,x_2$$

$$s = x_1$$

Système avec quatre constantes cinétiques (a_{ij}).

$$\frac{dx_1}{dt} = A_{11}\,x_1 + A_{12}\,x_2 + u$$

$$\frac{dx_2}{dt} = A_{21}\,x_1 + A_{22}\,x_2$$

$$s = x_1$$

modèle à quatre paramètres (A_{ij})

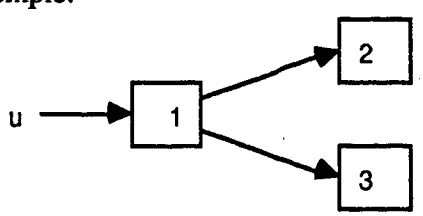
$$T(p) = \frac{X_1(p)}{U(p)} = \frac{A_{22}\,P}{p^2 - p(A_{11} + A_{22}) + A_{11}\,A_{22}}$$

fonction de transfert à trois paramètres A_{22}, $A_{11} + A_{22}$, $A_{11}\,A_{22}$

Donc *modèle non identifiable.*

NB: le système est commandable (si $a_{21} = 0$) et observable (si $a_{12} = 0$)

Autre exemple:



$$s = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} x$$

Système à trois constantes cinétiques

$$x^0 = \begin{bmatrix} A_{11} & 0 & 0 \\ A_{21} & 0 & 0 \\ A_{31} & 0 & 0 \end{bmatrix} \underline{x} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} u$$

modèle à trois paramètres

$$T1(p) = \frac{A_{21}\, p}{p^3 - A_{11}\, p^2} \qquad T2(p) = \frac{A_{31}\, p}{p^3 - A_{11}\, p^2}$$

Fonctions de transfert à trois paramètres (A_{11}, A_{31}, A_{21}). Donc modèle à trois paramètres identifiables alors que le système n'est pas commandable.

Pour étudier l'identifiabilité des modèle non linéaires, il existe peu de résultats théoriques, cependant, elle peut être en partie approchée par l'analyse de sensibilité.

Soit un modèle défini par:

$$(1)\quad \begin{aligned} x_{k+1} &= f(x_k, u_k, p) \quad \text{----------} \quad \text{état} \\ s_k &= g(x_k) \quad \text{----------} \quad \text{sortie} \end{aligned}$$

en linéarisant on a :

$$(2)\quad \Delta x_{k+1} = \frac{\partial f}{\partial x}\Delta x_k + \frac{\partial f}{\partial p}\Delta p = A_k\, \Delta x_k + B_k\, \Delta p$$

$$\Delta s_k = \frac{\partial g}{\partial x}\Delta x_k = c_k\, \Delta x_k$$

on peut ainsi exprimer Δs en fonction Δp.

$$\Delta s_{k+1} = C_{k+1}\left[\sum_{j=0}^{k-1}\left(\prod_{i=j+1}^{k} A_i\right) B_j + B_k\right]\Delta p$$

$$(3)\quad \Delta s_{k+1} = S_{k+1}\, \Delta p$$

S_{k+1} = matrice de sensibilité = $\frac{\Delta s}{\Delta p}$ de dimension n x r (n = nombre de variables de sortie et r = nombre de paramètres).

Le modèle linéarisé (2) sera identifiable si l'équation (3) a une solution unique, c'est-à-dire si la matrice S est de rang plein et donc a des colonnes linéairement indépendantes.

- deux colonnes proportionnelles signifient que deux paramètres ont des effets identiques sur la sortie, on conçoit alors qu'on ne puisse pas les distinguer, donc pas les identifier à partir de la sortie.
- une colonne nulle indique qu'un paramètre n'a aucun effet sur la sortie, il n'est donc pas identifiable à partir de cette sortie.

Ces propriétés d'identifiabilité ne sont valables que *localement*, c'est-à-dire que dans le domaine où le modèle linéarisé (2) est valable.

En pratique, lorsqu'un modèle n'est pas identifiable, il convient:

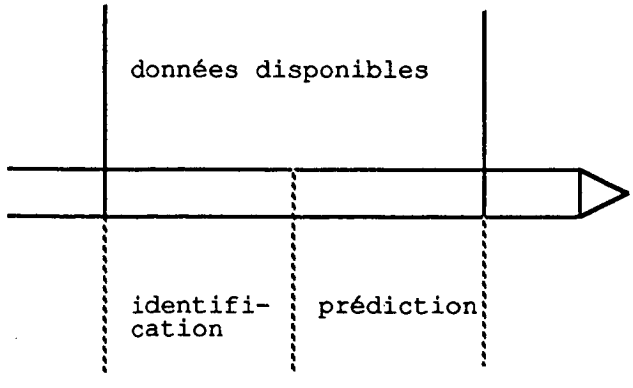
- soit de modifier les conditions expérimentales, par exemple les entrées u_k (S_k dépend de u_k) ou de changer de sortie, ou de considérer d'autres instants de mesures des mêmes sorties...
- soit de modifier le modèle.

En conclusion, l'identification reste une étape délicate qui comporte des aspects théoriques, expérimentaux et numériques:
* théoriques pour l'étude de l'identifiabilité théorique et de l'estimation (propriétés des estimateurs);
* expérimentaux pour la conception d'une expérimentation bien adaptée (choix des variables, des instants de mesures optimaux...);
* numériques pour l'optimisation du critère d'écart.

VII - Validation du modèle

Un bon ajustement entre données expérimentales et simulées à l'issue de l'identification n'est pas suffisant pour considérer un modèle comme prêt pour l'application. Il doit être validé. A notre connaissance, il n'existe pas de méthode universelle de validation, cependant divers tests peuvent être mis en œuvre pour apprécier la validité du modèle:

- *des tests expérimentaux*, généralement spécifiques à chaque étude et qui consistent à :
  * s'assurer que les valeurs estimées des paramètres qui ont une signification biologique sont d'ordre de grandeur acceptable;
  * analyser le comportement du modèle dans des conditions expérimentales voisines de celles qui ont servi à l'identification;
  * tester les capacités prédictives du modèles.



- *des tests statistiques* dérivés de la théorie de l'estimation. Ils dépendent de l'estimateur utilisé dans l'identification: par exemple, le test F qui vérifie si l'erreur dûe à l'approximation des mesures par le modèle est négligeable devant la dispersion des mesures.

VIII - Application du modèle

L'utilisation du modèle doit correspondre exactement à l'objectif que l'on s'était fixé au départ. Il est important de souligner que les résultats de l'application dépendent, entre autre, de la qualité et de la validité du modèle utilisé. Ainsi les performances d'une commande optimale dépendent non seulement de la méthode de commande mais aussi de la validité du modèle utilisé pour la calculer. Un exemple typique sera présenté. Il s'agit d'un problème de commande de procédé de fermentation utilisé pour la biosynthèse d'antibiotique. L'objectif était de maximiser la production d'antibiotique en agissant sur la température et le pH du fermenteur. Un modèle a été construit, une commande déterminée et une augmentation de 10 % de la production étaient escomptées. Expérimentalement, aucune augmentation n'a été obtenue car le modèle utilisé n'était pas valable durant la phase finale de fermentation.

Conclusion

La méthodologie de la modélisation qui vient d'être présentée est un outil de travail intéressant pour celui qui a à élaborer un modèle mathématique, car elle s'appuie sur une expérience et sur de nombreuses études de cas qui ont permis de répertorier les difficultés et les écueils. Elle est une sorte de "guide opérateur" pour le modélisateur.

Cependant , la mise en œuvre de cette méthodologie reste délicate car les choix qui doivent être faits au niveau des diverses étapes font appel à des compétences variées difficiles à réunir (mathématique, biologie, informatique, automatique...) Aussi pour faciliter sa mise en œuvre, cette méthodologie devrait être accompagnée d'outils d'aide à la modélisation. Certains existent déjà actuellement, par exemple en simulation et identification mais ils se présentent généralement comme un recueil de méthodes ou programmes, parmi lesquels l'utilisateur doit choisir le mieux adapté à son problème. Or bien souvent ce choix est difficile car l'utilisateur n'a pas la compétence requise. Il conviendrait donc de développer des systèmes d'aide à la modélisation qui incluent une certaine connaissance, ce qui devrait être possible avec les techniques d'intelligence artificielle (projet EDORA).

Bibliographie

BOISVIEUX J.F., (1977). "Modélisation et commande des processus biologiques: Aspects théoriques et mise en œuvre. Thèse d'état . Paris VI.

BOURDAUD D., (1974). "Contribution à l'identification et à l'optimisation de procédés de fermentation". Thèse de 3ème cycle. Grenoble.

CAMUS J.L., (1980). "Modélisation et commande de l'apport en substrat du procédé de biosynthèse de l'érythromycine". Thèse de docteur ingénieur. Grenoble.

CHERUY A., (1987). "Analyse cinétique d'un biosystème". Thèse d'état. Grenoble.

CHERUY A., DURAND A., (1979). "Optimisation of erythromycin biosynthesis by controlling pH and temperature = theoretical and practical aspects application. Biotech. Bioeng. n° 9, 303-320.

CHERUY A., GAUTIER C., PAVÉ A., (1981). "La notion de système dans les sciences contemporaines""Analyse de systèmes biologiques: aspects méthodologiques liés à la modélisation" . J. Lesourne Ed. tome I, 73-153. .

CHEVALLIER F., (1984). "Systèmes et modèles" Ed. CNRS.

ENDRENY L., (1981). "Kinetic data analysis" Plenum Press.

FREIN Y., (1983). "Modélisation par systèmes à compartiments: application au métabolisme du cholestérol". Thèse de docteur ingénieur . Grenoble.

GENTIL S., (1981). "Modélisation en écologie: méthodologie et application aux écosystèmes lacustres". Thèse d'état. Grenoble.

GENTIL S., CHERUY A., (1981). "Modélisation des systèmes biologiques et écologiques: Pourquoi et Comment ? Colloque Rythmes Oscillations et Modèles. Pau.

GLEASON-GARCIA E., (1978)."Modélisation d'une station d'épuration biologique des eaux à l'oxygène". Thèse de docteur ingénieur. Grenoble.

HOLMBERG A., RANTA J., (1982). "Procedures for parametres and state estimation of microbial growth processes". Automatica vol. n°18, 181-193.

ILIADIS A., (1980). "Modélisation du système coagulolytique sanguin - application au diagnostic". Thèse d'état. Grenoble.

LEBRETON J.D., MILLIER C., (1982). "Modèles dynamiques déterministes en biologie". Masson.

LUKASIK A., (1974) "Sur l'identification de procédés de fermentation" Thèse de 3ème cycle. Grenoble.

PAVÉ A., (1980). "Contribution à la théorie et à la pratique des modèles mathématiques pour l'analyse dynamique des sytsèmes biologiques". Thèse d'état. Lyon.

SPRIET, VANSTEENKISTE, (1982). "Computer aided modelling and simulation" Academic Press.

VIALAS Ch., (1984). "Modélisation et contribution à la conception d'un procédé biotechnologique". Thèse docteur ingénieur. Grenoble.

WALTER E., (1982). "Identifiability of state models" Lectures notes in biomathematics : Springer Verlag ed.

ANNEXE

Analyse de données biologiques expérimentales en vue d'une modélisation

Les données biologiques à analyser concernent le système cholestérol qui peut être vu comme un système à deux entrées (absortion et synthèse) et à deux sorties (excrétion et transformation en particulier en acides biliaires), et dont nous nous intéressons au fonctionnement interne, notamment aux échanges entre organes et aux tranformations entre formes libre et estérifiée qui sont celles sous lesquelles se trouve le cholestérol dans l'organisme. Les organes considérés ici sont le foie, l'intestin, le plasma, les hématies et la carcasse qui regroupe tout le reste de l'organisme. L'expérimentation disponible est faite sur le rat; le protocole expérimental comprend quatre expériences différentes (A, H, P, S) utilisant des traceurs radioactifs et consistant à suivre dans le temps l'évolution du cholestérol marqué dans les différents organes. Pour ce faire, pour chaque temps expérimental, des rats doivent être sacrifiés, leurs organes prelévés et le cholestérol extrait. Ainsi chaque mesure correspond à un lot d'animaux différents.

* Expérience A (Absorption): les rats ingèrent au temps t = 0 une nourriture dont le cholestérol est marqué.
* Expérience H (Hématies): on injecte en début d'expérience des hématies dont le cholestérol est marqué.
* Expérience P (Plasma): on injecte à t = 0 du plasma dont le cholestérol est marqué.
* Expérience S (Synthése): on injecte un précurseur radioactif du cholestérol (acétate).

Ainsi, les résultats expérimentaux pour chaque expérience peuvent être présentés sous forme de tableaux donnant pour chaque temps expérimental la mesure de la radioactivité spécifique du cholestérol libre et estérifié dans chaque organe. Au niveau du plasma, la mesure est affinée par le dosage du cholestérol dans quatre unités lipoprotéiques (HDL, LDL, VLDL et chylomicrons).

Le tableau 1 est un exemple de présentation des résultats expérimentaux, et la figure 1 la structure a priori du système.

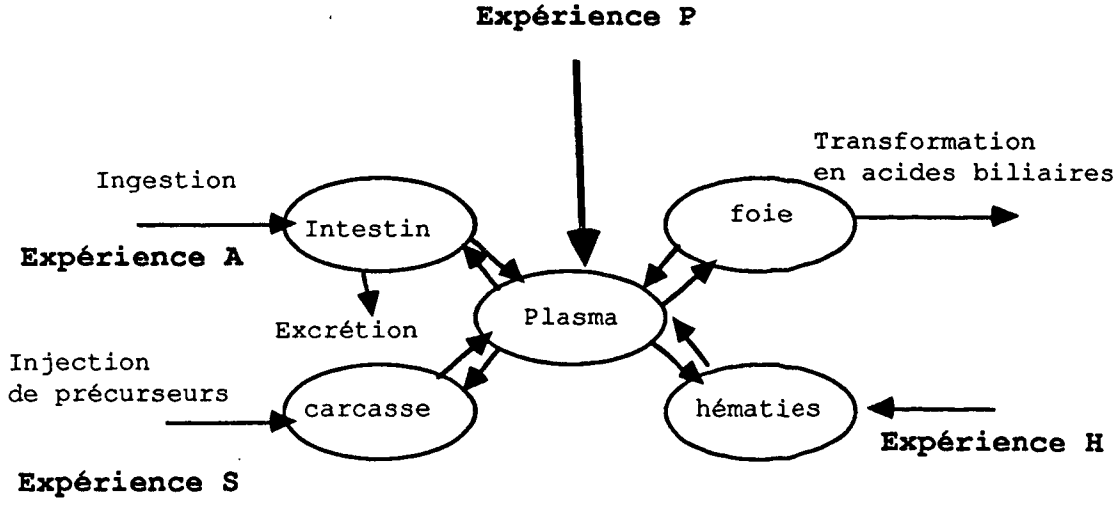


Figure 1- Structures a priori du système

Temps expérimental		3	3	5	5	10	10	14	14	23	27	27	50	50
HDL	CHL	654	651	1774	1455	2782	3758	4377	4334	4411	6581	6456	4660	4142
	CHE	234	271	910	1045	2154	2917	3624	3722	4275	5035	4602	5178	4349
LDL	CHL	729	553	2456	2001	3186	4375	4754	4895	3729	5983	5151	4556	4246
	CHE	126	119	318	364	987	1234	1836	1734	3093	3041	3365	4869	4453
VLDL	CHL	1355	1140	2456	2774	3724	5048	6542	6680	4684	3191	7418	4556	3995
	CHE	2570	2388	8413	9096	8974	10994	13462	12646	8777	4288	8379	6627	4504
CHYLOMICRONS	CHL	4768	5265	4457	5912	5833	5946	6684	6680	4866	17550	13462	4504	4090
	CHE	12994	18347	9732	10596	5429	5048	5742	6119	3420	8276	7418	6213	11080
HEMATIES	CHL	187	190	682	773	1795	3950	3389	3467	3820	5185	4876	4456	4400
FOIE	CHL	776	706	2137	2183	3814	4936	5931	5915	4275	5235	5082	4349	4090
	CHE	916	412	1364	1364	2782	2131	3907	3569	4548	4786	4739	4763	4608
INTESTIN	CHL	9815	7925	13280	11733	15346	17276	17227	18357	7113	4687	6456	3365	3210
	CHE	14770	13353	15735	13825	4622	5160	21557	10600	7067	3869	7349	3314	3107
CARCASSE	CHL	126	146	116	116	276	344	403	437	638	878	1209	950	950
	CHE	35	41	50	50	176	220	108	117	202	267	368	315	315

Tableau 1: Exemple de résultats expérimentaux "EXPERIENCE ABSORPTION"

CHL: Cholestérol libre (mesures exprimées en radioactivité spécifique)
CHE: Cholestérol estérifié (dpm/mg)

Analyse des données

L'objet des méthodes d'analyse des données est de résumer l'essentiel de l'information contenue dans des tableaux de données types [variables, individus]. Dans notre cas, les variables correspondent aux temps expérimentaux et les individus au cholestérol libre et estérifié dans les différents organes.

Parmi les méthodes disponibles, l'analyse des correspondances s'est révélée la mieux adaptée. Cette méthode consiste à résumer le plus fidèlement possible l'ensemble des données expérimentales par des projections dans des sous-espaces, la fidélité de la représentation obtenue peut être mesurée par l'importance des facteurs (axes de l'ellipsoide d'inertie de l'ensemble des points expérimentaux pondérés).

Sa mise en œuvre sur nos données a montré que les deux premiers facteurs expriment plus de 80 % de la variabilité, ainsi une projection dans le plan des deux axes principaux fournira une image assez fidèle.

Le premier facteur (60 % à 78 % de la variabilité suivant les expériences) exprime tout simplement la chronologie des temps expérimentaux (résultat classique dans le cas de processus temporels).

L'examen des cartes factorielles dans le plan permet de repérer les positions relatives des individus et des temps expérimentaux.

Celle relative à l'expérience "hématies" est particulièrement facile à interpréter. Les hématies, qui ne sont porteuses que de cholestérol libre, devancent de loin tous les autres individus. Les échanges de cholestérol entre les hématies et les lipoprotéines plasmatiques expliquent que, dans un deuxième temps, viennent les individus correspondant au cholestérol libre de ces lipoprotéines.

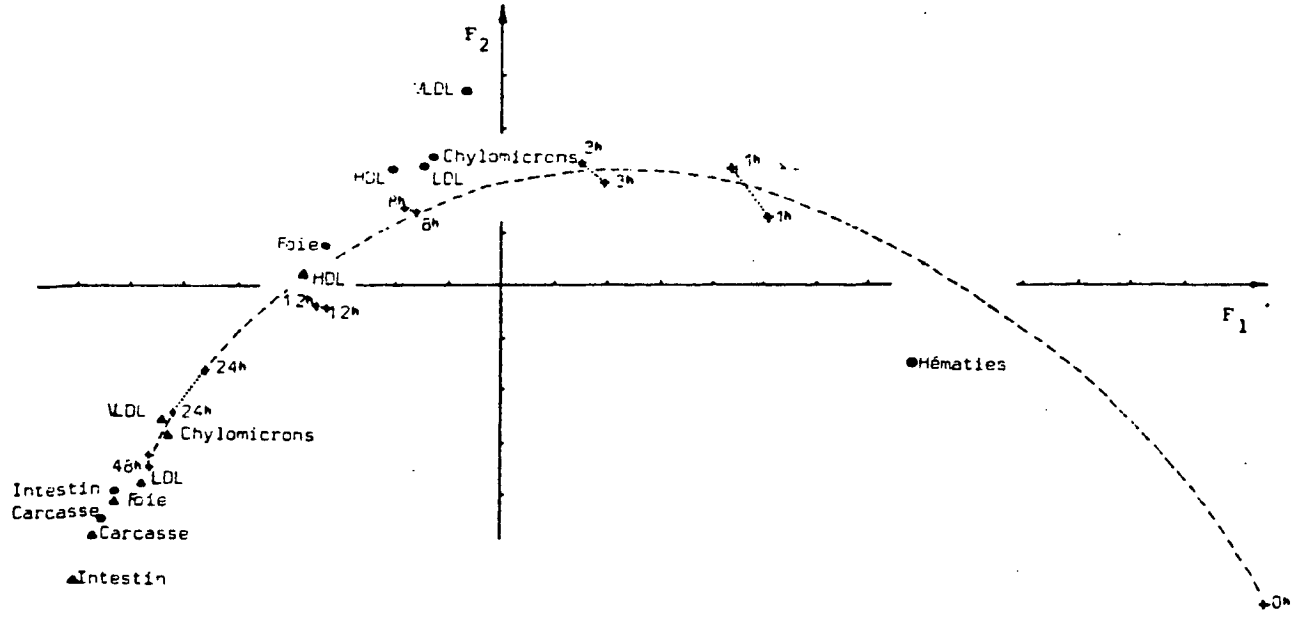


Figure 2 - Carte factorielle de l'expérience "Hématies"

Le foie, étant l'objet d'un intense échange de cholestérol libre avec les lipoprotéines, apparaît peu après. Par contre, les individus "cholestérol libre de l'intestin et de la carcasse sont très décalés (les échanges de cholestérol libre sont moins rapides à leur niveau). Ils font partie d'un troisième et dernier groupe qui comprend la plupart des individus "Cholestérol estérifié". Il est remarquable de constater que dans tous les cas, la forme estérifiée d'un substrat n'intervient que secondairement à sa forme libre. De plus, le rôle charnière joué par les HDL est très clair: bien que leur cholestérol libre soit le dernier marqué de toutes les lipoprotéines, leur cholestérol estérifié devance tous les autres individus "Cholestérol estérifié". Il apparaîtrait que, dans le plasma, les HDL seraient responsables de l'estérification du Cholestérol.

Pour analyser les résultats relatifs à l'expérience "Absorption", le même plan peut être suivi. Pour les temps courts, apparaissent les individus qui jouent un rôle dans la commande isotopique du système. Ce sont le cholestérol (libre et estérifié) de l'intestin, organe de transit du cholestérol alimentaire absorbé, et les chylomicrons puis les VLDL qui le véhiculent sous forme estérifiée.

Cependant l'expérience "Absorption" n'est sûrement pas la plus appropriée pour clarifier le métabolisme dans son ensemble. L'analyse détaille simplement le métabolisme du cholestérol à sont entrée dans l'organisme.

Pour l'analyse des données relatives à l'expérience "Plasma", on retiendra que le premier groupe d'individus à supporter l'essentiel de la variabilité correspond à toutes les lipoprotéines, à égalité d'importance pour les deux formes chimiques du cholestérol. Cette observation nous a conduit à regrouper toutes les lipoprotéines et à ne considérer qu'une seule variable pour le plasma dans le modèle des échanges de cholestérol.

Enfin, la carte factorielle correspondant à l'expérience "synthèse" fait ressortir le rôle primordial de l'instestin, résultat original pressenti par les biologistes mais qui avait été longtemps contesté.

Par ailleurs, l'analyse de données a permis de montrer que la variabilité inter lot est suffisamment réduite pour ne pas nuire à l'utilisation et à l'interprétation des résultats. Enfin, nous avons observé que les temps longs apportent peu d'observation ce qui conduit à ne pas en tenir compte au niveau de la modélisation (surtout dans le cas des échanges) et d'autre part à alléger le dispositif expérimental.

En conclusion, on peut considérer que l'apport de l'analyse des données a été relativement limité, ce qui peut s'expliquer par la connaissance que l'on avait, a priori, du système: les résultats avaient été "bien" analysés "à la main". Cependant, l'emploi de ce type de méthode a permis de rassembler, d'ordonner les connaissances, voire de conforter certaines hypothèses (par exemple l'intestin comme lieu essentiel de synthèse). Cet aspect présentation n'est déjà pas négligeable.

Références

ANFREVILLE R., (1983). Contribution à la modèlisation du système chez le rat. Thèse de 3ème cycle. Lyon.

CHEVALLIER F., (1977). Rat cholesterol as a dynamic system. Ed.par Boulanger P., Jayle M.-F., Roche J..Exposés annuels de Biochimie Médicale. Masson Publish, Paris, 87-113.

MAGOT Th., (1985). Contribution à la modélisation du système cholestérol du rat. Thèse d'état. Orsay.

INTERPRETATION ET CONSTRUCTION DE MODELES DE LA DYNAMIQUE DES POPULATIONS A L'AIDE DE SCHEMAS FONCTIONNELS

Alain PAVÉ
Laboratoire de Biométrie et de Biologie des Populations
UA CNRS 243
Université Claude Bernard - Lyon 1
69622 Villeurbanne Cedex

1. Introduction

Ce travail s'inscrit dans le projet EDORA d'élaboration d'un système "intelligent" d'aide à la modélisation en Biologie et en Ecologie. Il a été effectué dans le but d'organiser une base de connaissances alliant **modèles**, en particulier leur(s) formulation(s) mathématique(s) et leurs propriétés intéressantes, **interprétations** de ceux-ci en termes biologiques, et **relations** entre modèles tant du point de vue mathématique que de l'interprétation biologique. On s'intéresse également au problème de la **construction** de nouveaux modèles. Le champ choisi est celui des modèles différentiels de la dynamique des populations.

Pour l'interprétation et la construction des modèles, l'utilisation de **représentations schématiques** est commode (e.g. la représentation en "boîtes et flèches" des systèmes à compartiments, la notation symbolique des réactions chimiques, les "bond graphs"...). Il s'agit d'un formalisme intermédaire entre les hypothèses discursives, concernant la structure et/ou le fonctionnement d'un système, et la formulation mathématique opérationnelle. Quand ces représentations sont plus orientées vers l'aspect fonctionnel, donc en relation avec la dynamique des systèmes, on parle de **schémas fonctionnels**. L'intérêt de ces représentations réside dans leur simplicité, leur pouvoir de description, et la possibilité, dans un certain nombre de cas, de traduire les schémas en expressions mathématiques, traduction qui peut être automatisée. Inversement on peut proposer un procédé d'inférence conduisant, sous certaines contraintes, à proposer un, ou plusieurs, schéma(s) fonctionnel(s) associé(s) à un modèle (i.e. tel que si on applique à ce schéma l'algorithme de traduction on trouve une expression équivalente à l'expression mathématique typique du modèle).

Dans le cadre choisi, une formulation "**type chimique**" s'est avérée relativement bien adaptée pour l'interprétation et la construction de modèles différentiels de la dynamique des populations. Cette formulation permet aussi d'étudier les relations entre modèles conduisant à une tentative de classification de ces modèles. Enfin, l'aspect d'interprétation permet de distinguer le niveau phénoménologique ou "superficiel" (celui de l'observation, ce que décrit le modèle) du niveau explicatif ou "profond" exprimé en termes de processus biologiques.

1.1. Schémas fonctionnels

Nous avons discuté à plusieurs occasions de l'utilisation de représentations schématiques pour l'interprétation et la construction de modèles mathématiques (e.g. **Pavé** et **Pagnotte**, 1977, **Pavé**, 1980, **Pavé** et **Rechenmann**, 1986, **Pavé**, 1987). Ces représentations schématiques sont des intermédiaires entre un ensemble d'hypothèses discursives émises sur la structure et/ou le fonctionnement d'un système, et un modèle

mathématique. De telles représentations sont connues et utilisées dans de nombreux domaines. On peut citer quelques exemples:

- les diagrammes en "boîtes et flèches" en analyse compartimentale,
- la notation symbolique des réactions chimiques,
- les diagrammes en blocs en ingéniérie,
- les "bonds graphs" en mécanique et dans d'autres secteurs,
- les diagrammes de Forrester dans le domaine socio-économique.

Elles sont caractérisées par:

- l'utilisation d'un ensemble limité de symboles, et d'une syntaxe élémentaire. Dans ce sens de sont de véritables **langages de description**..
- l'association avec une classe d'objets mathématiques (par exemple, des systèmes différentiels linéaires pour l'analyse compartimentale linéaire);
- l'existence d'un algorithme de traduction des représentations schématiques en expressions mathématiques. Certains logiciels de simulation offrent cette facilité à l'utilisateur (par exemple, COSMOS (**Hamrouni**, 1979), et plus récemment STELLA (**Richmond**, 1985)).

Les symboles utilisés et leurs associations ont une signification précise: ils correspondent à des variables, des paramètres, interprétables en termes biologiques (taille d'une population, concentration d'un médicament dans le sang, taux de reproduction...), ou à des relations entre ces symboles (flux, interactions)... Ces représentations sont, entre autres, de bons moyen de décrire les aspects fonctionnels d'un système, aussi parle-t-on de **schémas fonctionnels**.

1.2. Représentation schématique et formulation mathématique

Le processus de passage de la représentation schématique au modèle mathématique est généralement bien connu (aspect **construction**), par contre la démarche inverse a été beaucoup moins étudiée. Elle ne semble pourtant pas dénuée d'intérêt: on peut s'interroger sur l'**interprétation** d'expressions mathématiques connues dans un domaine donné comme bons descripteurs de phénomènes observés.

En outre une base de connaissance sur les modèles mathématiques doit contenir non seulement des aspects mathématiques mais également des liens avec une connaissance biologique. L'interprétation en termes de schémas fonctionnels est une des voies permettant d'intégrer ce type de connaissance.

En l'occurrence, je me suis plus particulièrement intéressé à ce point de vue dans un cas particulier: celui des **modèles** classiques de la **dynamique des populations** (dits de **"Lotka-Volterra"**), exprimés sous formes différentielles et intégro-différentielles. La représentation schématique choisie est voisine de celle utilisée pour représenter les réactions chimiques, on conviendra de l'appeler "**représentation type chimique**". Celle-ci a été choisie après avoir remarqué la parenté entre les modèles de la cinétique chimique et ceux de la dynamique des populations. **Garfinkel** (1962) avait déjà souligné l'intérêt de ce formalisme pour la construction de modèles en écologie. Il est clair qu'il ne s'agit pas de l'outil miracle mais d'une première approche dont peut-être le principe pourra être repris ou étendu à d'autres situations avec d'autres formalismes.

Dans une première partie on rappelle les principes de l'algorithme de traduction du schéma fonctionnel en expression mathématique, ainsi que les bases du processus d'inférence du schéma fonctionnel à partir d'une formulation mathématique. On introduit une possibilité nouvelle par rapport aux travaux déjà publiés: l'introduction de termes de saturation de type "Monod" (ou michaëlien). Ensuite une étude des principaux modèles de la dynamique des populations, illustrés par des **exemples précis** proposés en fin de texte, nous conduit à proposer une classification de ces modèles fondée sur les schémas fonctionnels. Cette classification recouvre à quelques détails près celle qui a été présentée dans deux articles récents (**Pavé et Rechenmann**, 1986, **Pavé**, 1987). On peut enfin

situer le niveau d'intervention de la schématisation dans le processus de modélisation (figure 1).

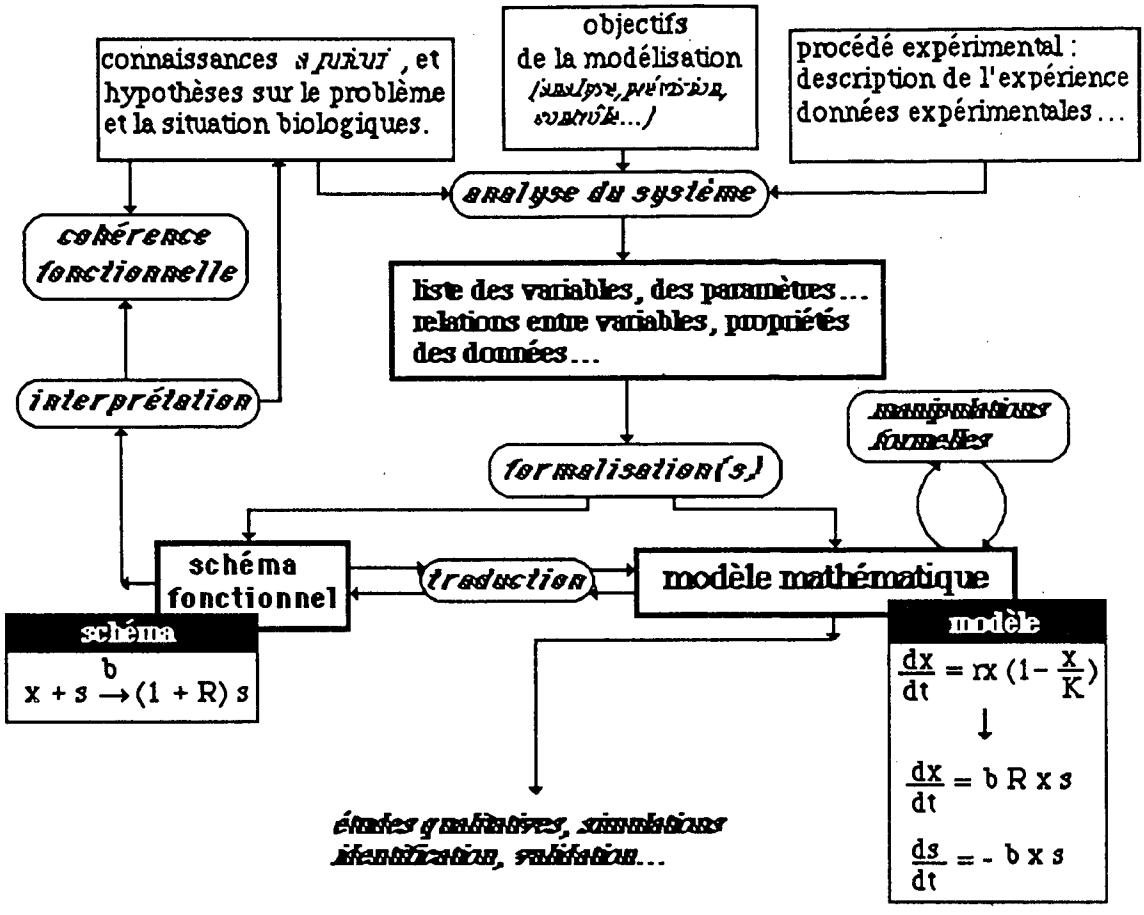


Figure 1 - Dans le diagramme illustrant la démarche de modélisation (cf. A. Chéruy, même volume) on peut situer l'intervention des schémas fonctionnels que ce soit au niveau de la construction du modèle ou à celui de l'interprétation. Les illustrations placées dans les fenêtres "modèle" et "schéma" sont relatives au modèle logistique, et correspondent à une interprétation de ce modèle (cf. 2.2.).

2. Représentation type chimique et modèles différentiels multilinéaires

Les relations entre les systèmes d'équations différentielles ordinaires multilinéaires de la cinétique chimique et le formalisme utilisé pour représenter les réactions est connu depuis longtemps, on pourra en trouver un exposé actuel dans l'ouvrage d'**Emanuel et Knorre** (1975).

Cependant, il faut retenir les travaux de **Garfinkel** (1962), qui s'est intéressé tout particulièrement au problème de la programmation de l'algorithme de traduction (**Garfinkel**, 1961). En ce qui nous concerne nous avons proposé une reformulation du problème en 1977 (**Pavé et Pagnotte**, 1977), publiée d'ailleurs indépendamment et sous une forme voisine par **Vidal** en 1978. Cette nouvelle formulation permet de mieux aborder le problème d'inférence du schéma fonctionnel à partir de l'expression mathématique.

2.1 - Principaux éléments sur l'algorithme de traduction (Pavé et Pagnotte, 1977; Pavé, 1980; Pavé et Rechenmann, 1986).

Un ensemble de réactions type chimique s'écrit:

$$S = \{ n_{i1}X_1 +...+ n_{ij}X_j +...+ n_{iq}X_q \xrightarrow{a_i} m_{i1}X_1 +...+ m_{ij}X_j +...+ m_{iq}X_q \}$$

pour $1 \leq i \leq r$ (r est le nombre de réactions simultanées);
les n_{ij} sont des entiers positifs, ou nuls (dans ce cas le terme correspondant est omis);
les m_{ij} sont des réels positifs, ou nuls (dans ce cas le terme correspondant est aussi omis);

les X_j sont les symboles correspondants aux variables d'état liées aux espèces qui interagissent (ou plus brièvement les espèces elles-mêmes), dont la dynamique est représentée par le système différentiel:

$$\frac{dx_i}{dt} = \sum_{i=1}^{r} a_i \, (m_{ij} - n_{ij}) \prod_{k=1}^{q} x_k^{\,n_{ik}}$$

Cette notation s'interprète comme l'interaction des termes de la partie gauche (dans les proportions n_{ij}) qui produit les termes de la partie droite (dans les proportions m_{ij}), les constantes de vitesse des réactions sont symbolisées par les a_i. Cette formulation est fondée sur une hypothèse d'interaction entre les espèces régie par une loi du type "action de masse" (**Garfinkel**, 1962). Par exemple, prenons deux espèces x et y qui se nourrissent d'un même substrat, on pourra représenter le système, en milieu isolé et limité en substrat, par le schéma fonctionnel:

$$x + s \xrightarrow{a_1} (1 + R_1)\, x$$
$$y + s \xrightarrow{a_2} (1 + R_2)\, y$$

auquel on fait correspondre le système différentiel:

$$\frac{dx}{dt} = a_1 R_1 \, x \, s$$
$$\frac{dy}{dt} = a_2 R_2 \, y \, s$$
$$\frac{ds}{dt} = - a_1 \, x \, s - a_2 \, y \, s$$
$$x(0) = x_o,\ y(0) = y_o,\ s(0) = s_o$$

2.1.1. Notation matricielle

On conviendra de représenter un système de r réactions en utilisant les notations matricielles suivantes:

-pour le schéma fonctionnel

$$D_d X \xrightarrow{A} D_g X$$

X est la matrice unicolonne des espèces, D_g est la matrice r x q des coefficients des espèces apparaissant dans les parties gauches des réactions (les n_{ij}), D_d est la matrice des coefficients des espèces apparaissant dans les parties droites des réactions (les m_{ij}).

- pour le système différentiel

$$\frac{dX}{dt} = D\,A\,V$$

avec $D = D_d^T - D_g^T$ où D_d^T (resp D_g^T) représente la transposée de D_d (resp. D_g), i.e. $d_{ki} = m_{ki} - n_{ki}$), V est la matrice colonne dont le $i^{ème}$ terme est $\prod_{k=1}^{q} x_k^{nik}$, enfin A est la matrice diagonale des constantes de vitesse a_i.

-Reprenons l'exemple ci-dessus, on a

$$X = \begin{pmatrix} x \\ y \\ s \end{pmatrix},\ D_g = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix},\ D_d = \begin{pmatrix} 1+R_1 & 0 & 0 \\ 0 & 1+R_2 & 0 \end{pmatrix}$$

$$D = \begin{pmatrix} R_1 & 0 \\ 0 & R_2 \\ -1 & -1 \end{pmatrix},\ A = \begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix} \text{ et } V = \begin{pmatrix} x & s \\ y & s \end{pmatrix}$$

2.1.2. Ecriture du système différentiel à partir d'un schéma fonctionnel

A partir d'un système de r réactions constituant le schéma fonctionnel on peut aisément générer de façon **symbolique**: la matrice V à partir des parties gauches des réactions, la matrice A des constantes de vitesse puis la matrice D, et enfin développer le système différentiel correspondant en appliquant les règles usuelles du produit matriciel sur les symboles. Certaines simplifications sont possibles notamment en examinant le rang de la matrice D, ou en éliminant les termes qui interviennent uniquement en partie droite (qui n'influent donc pas dans la dynamique du système).

2.1.3. Inférence d'un schéma fonctionnel

Il est clair que si on dispose d'un système différentiel multilinéaire sous la forme proposée ci-dessus le problème est relativement simple à résoudre. Il suffit de générer la matrice V, la matrice D_g (à partir des exposants des variables d'état), la matrice A, la matrice D et finalement la matrice D_d ($D_d = D^T + D_g$).

On considèrera que la **condition d'existence** d'un schéma fonctionnel type chimique est que les n_{ij} (termes de D_g) doivent être des entiers naturels et les m_{ij} (termes de D_d) des réels positifs ou nuls (**Pavé**, 1980).

Le problème est ainsi apparemment simple à résoudre, cependant on peut faire les remarques suivantes:

(i) souvent à partir de l'expression différentielle on ne peut pas distinguer les matrices D et A mais obtenir seulement leur produit, il y a alors lieu de prendre certaines conventions (c'est d'ailleurs en partie le sens du choix du domaine des valeurs admises pour les m_{ij} , i.e. $\mathbf{R^+}$);

(ii) la représentation d'un système ouvert, dans le cas d'entrées à vitesse constante et de sorties proportionnelles à l'état x (pouvant aussi représenter un processus de mortalité), on peut convenir de la notation suivante:

$$1 \xrightarrow{u} x$$
$$x \xrightarrow{k} 0$$

(iii) on peut proposer des systèmes différentiels équivalents à un système donné (par exemple, par une manipulation algébrique, un changement de variable, ou en plongeant le système dans un espace de plus grande dimension), d'où des **schémas fonctionnels différents**, donc des **interprétations éventuellement différentes.**

Il n'est pas question de répondre ici dans les détails, des éléments peuvent être trouvés dans les articles publiés récemment (**Pavé et Rechenmann**, 1986, **Pavé**, 1987). Je me limiterai d'abord à discuter du point (iii) à travers un exemple: celui du modèle logistique.

2.2 - Exemple du modèle logistique

Ce modèle bien connu, et l'un des plus simples, de la dynamique des populations. On le trouve le plus souvent dans la littérature biologique sous la forme:

$$\frac{dx}{dt} = r\,x\left(1 - \frac{x}{K}\right)$$

Il est possible d'écrire des équations équivalentes. A partir de chacune d'elles on peut tenter de proposer des schémas fonctionnels (figure 2), et d'interpréter ces schémas fonctionnels en termes biologiques. Ainsi:

(S1) est interprétable en terme de *croissance limitée par un processus de compétition intraspécifique (voire prédation intraspécifique).*

(S2) peut s'interpréter comme *la croissance d'une biomasse* x *dans un milieu limité en substrat* s.

(S3.1) et **(S3.2)** représentent la *croissance d'une population sur un milieu limité en substrat, la biomasse est soumise à une dégradation (ou processus de mortalité).* Pour **(S3.1)**, cette dégradation génère une quantité équivalente de substrat à celle consommée pour produire la biomasse. C'est une hypothèse peut être trop forte dans la mesure où l'on sait que les produits de dégradation ne sont, en général, pas réutilisables en totalité comme substrat (du moins pas directement), c'est-à-dire que p_1 est plus petit que $1/R_1$. Une façon d'améliorer cette représentation, pour la laisser compatible avec le modèle logistique, est de supposer que la biomasse (les individus d'une population) est capable "d'exploiter" le milieu pour produire du substrat nécessaire à sa croissance et à son maintien, le schéma fonctionnel **(S3.2)** tient compte de cette situation.

(S4.1) et **(S4.2)**, ces schémas décrivent *la croissance d'une biomasse* x *en présence d'un facteur de croissance de type catalyseur* s qui se dégrade spontanément (décroissance exponentielle de s décrite par la deuxième "réaction"). Si $n > 2$ il y a production de facteur de croissance par la biomasse x. Noter que pour $n=1$ on retrouve le schéma **(S2)**.

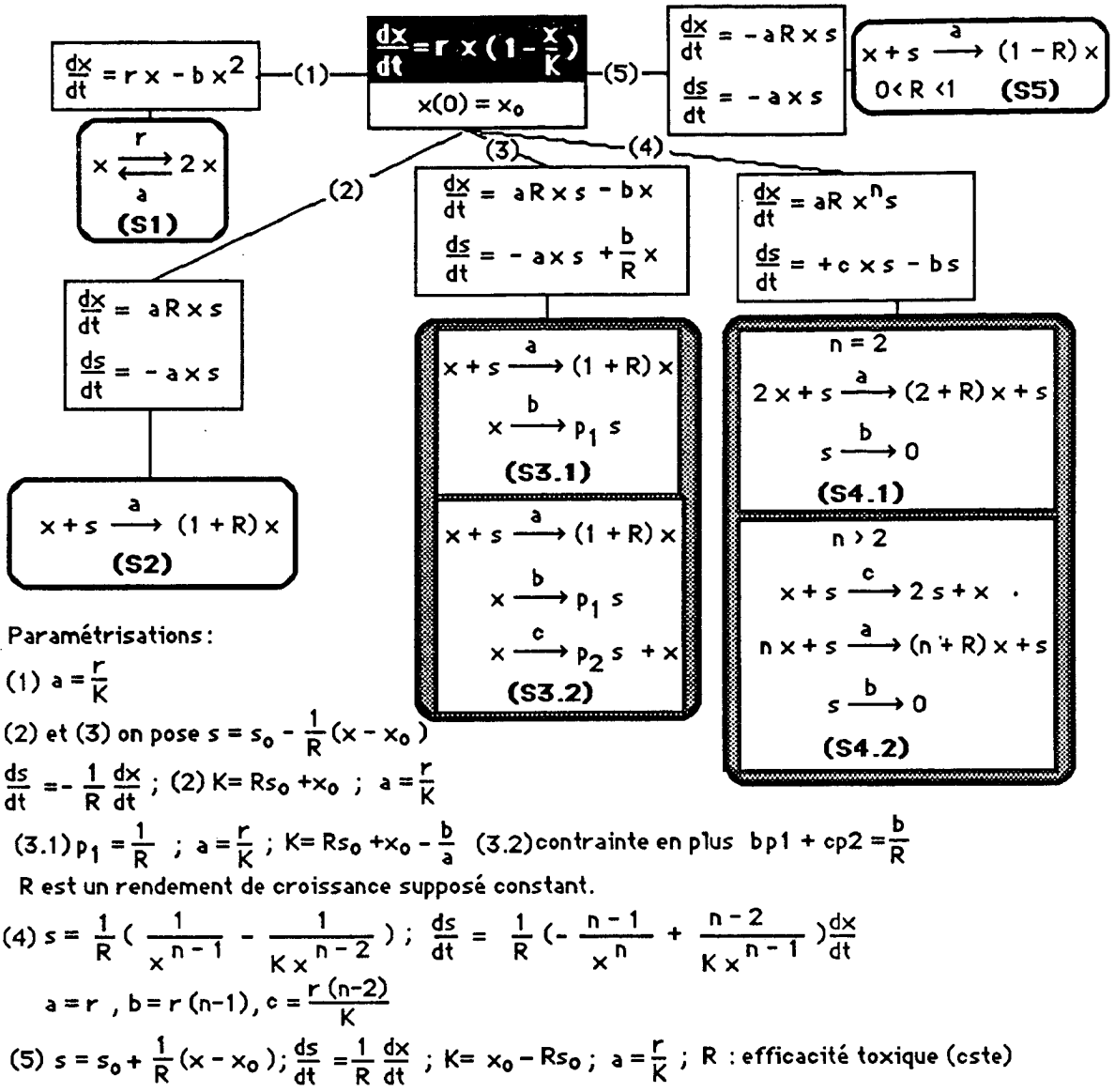


Figure 2 -. Le modèle logistique: différentes expressions équivalentes et schémas fonctionnels associés. L'interprétation biologique de ces différents schémas se trouve dans le texte.

2.3. Discussion

Cet exemple montre bien qu'on peut proposer **plusieurs schémas fonctionnels**, donc plusieurs interprétations, **pour un même modèle**. On peut noter que chacun de ces schémas correspond à une écriture mathématique précise, la non univocité provient donc des écritures équivalentes à ce au niveau: *formules équivalentes à une transformation algébrique près, mais à interprétations différentes* . En tout état de cause, le modèle logistique peut représenter des phénomènes de natures assez différentes, ce qui explique peut-être son bon pouvoir descriptif et sa large utilisation dans la littérature biologique. Inversement il faut se méfier de l'interprétation qu'on peut faire; en ce sens la représentation type chimique peut aider à vérifier la **cohérence** de l'utilisation du modèle en fonction du phénomène biologique observé (il s'agirait en quelque sorte d'une validation fonctionnelle). Par exemple, le modèle logistique peut décrire la croissance de populations d'organismes (les schémas S1, S2 et S3 présentent des mécanismes divers), il permet également de décrire la croissance de certains organismes

(les schémas S4 qui font intervenir un facteur de croissance seraient mieux adaptés à représenter une telle situation, on en rediscute pour le modèle de Gompertz).

2.4. Phénomènes de saturation

Les monômes algébriques proposés intervenant dans les vitesses des phénomènes supposent une stricte proportionnalité entre les variables d'état: par exemple la vitesse de croissance croit proportionnellement à la concentration ou à la quantité de substrat. On peut imaginer qu'il existe des possibilités de consommation limitées, ainsi Monod (1942) a-t-il proposé de remplacer ces termes multiplicatifs par des termes hyperboliques (de type michaëliens comme en cinétique enzymatique) pour traduire ces phénomènes de limitation ou de saturation. On peut aujourd'hui distinguer les cas suivants:

- limitation de la consommation et de l'assimilation d'un substrat pour la biomasse. Ceci peut se traduire dans l'expression mathématique par le remplacement de s par $\frac{s}{K+s}$. Le terme correspondant à la croissance limitée par le substrat devient:

pour la biomasse x: $\frac{a\,R\,x\,s}{K+s}$, pour le substrat s: $-\frac{a\,x\,s}{K+s}$

- limitation de l'accès au substrat, due à sa propre densité de la biomasse. Ce phénomène peut aussi être traduit par un terme hyperbolique mais ici relatif à la biomasse en remplaçant x par $\frac{x}{K'+x}$, alors le terme de croissance devient:

pour la biomasse x: $\frac{a\,R\,x\,s}{K'+x}$, pour le substrat s: $-\frac{a\,x\,s}{K'+x}$

- enfin on peut combiner ces deux phénomènes de saturation, on obtient alors, pour le même exemple:

pour la biomasse x: $\frac{a\,R\,x\,s}{(K+s)\,(K'+x)}$, pour le substrat s: $-\frac{a\,x\,s}{(K+s)\,(K'+x)}$

Ecriture des schémas fonctionnels intégrant des phénomènes de saturation

L'écriture classique telle qu'elle est présentée ci-dessus permet de générer simplement les équations de la dynamique du système. Les biochimistes utilisent une notation voisine, mais les notions de base "enzyme et substrat" ont un rôle asymétrique (l'enzyme est plutôt un catalyseur), alors qu'on vient de voir que les éléments d'une interaction quelconque peuvent parfaitement induire un phénomène de saturation. On peut proposer, par exemple, que si une espèce x_j induit ce phénomène relativement aux autres, dans la réaction i, le terme correspondant dans la partie droite de la réaction soit noté

$$\frac{n_j\,x_j}{K_{ij}}$$

cette notation a l'avantage de donner l'intuition du rôle limitatif par le terme de division.

Exemple:

reprenons la croissance "logistique" , le schéma correspondant est:

$$x + s \xrightarrow{a} (1 + R)\,x \quad , \quad \text{terme en x: } a\,R\,x\,s$$

s'il y a saturation par excès de substrat , on a:

$$x + \frac{s}{K} \xrightarrow{a} (1 + R)\,x \quad , \quad \text{terme en x: } \frac{a\,R\,x\,s}{K+s}$$

Ce schéma correspond au **modèle de Monod.**

On voit immédiatement comment s'écrirait les schémas correspondant à la saturation par la biomasse; écrivons la double saturation, on obtient:

$$\frac{x}{K'} + \frac{s}{K} \xrightarrow{a} (1+R)\,x \quad , \quad \text{terme en x: } \frac{a\ R\ x\ s}{(K+s)(K'+x)}$$

Les règles de traduction de cette nouvelle notation sont simples à établir par analogie avec celles correspondant à la notation classique. Seuls les termes de la matrice T sont modifiés: les variables x_j induisant une saturation dans la réaction i sont remplacées par l'expression:

$$\frac{x_j}{(K_{ij} + x_j)}$$

3. Schémas fonctionnels de modèles de la dynamique des populations

De la même façon que pour l'exemple du modèle logistique, on peut proposer des schémas fonctionnels de modèles de la dynamique des populations et d'en discuter les interprétations. Dans certains cas l'examen des schémas obtenus conduit à certaines modifications et/ou des extensions "raisonnables" sur le plan biologique, et ainsi à construire des modèles originaux. Enfin ces représentations fonctionnelles offrent un cadre à la fois pour l'analyse de la parenté entre modèles et pour la décomposition du **phénomène décrit** par le modèle en processus plus fins (**aspect explicatif**).

On se contentera ici de résumer les résultats déjà publiés par ailleurs concernant des modèles **sans saturation** (i.e. sans la présence d'un "terme de Monod") pour les cas à une et à deux dimensions.

3.1. Modèles à une population

Le modèle logistique

Ce modèle, que nous avons examiné, est certainement le modèle le plus connu en biologie. Il fut proposé au milieu du XIXème siècle par Verhulst (**Verhulst**, 1838) pour décrire la croissance de populations humaines (en l'occurrence la population de la Belgique). Son succès est certainement dû à la simplicité de la formulation, à l'interprétation des paramètres en termes biologiques, et à la grande diversité des situations que ce modèle peut décrire. Comme l'écrivait **Lotka** (1935): "it has been found to fit very acceptably a number of observed examples of population growth". Cette diversité est sans doute en partie explicable par les différents schémas fonctionnels qui peuvent générer ce modèle.

Enfin les notions de **stratégies r et K** peuvent se rediscuter, dans le cadre restreint que nous proposons, en termes de rendement de croissance (R), de vitesse de croissance (caractérisée par la constante $a = \frac{r}{K}$, de mortalité (caractérisée par la constante b) pour le schéma (S3), éventuellement de s_o (quantité totale de substrat disponible pour une population donnée).

L'interprétation (S4) faisant intervenir un facteur de croissance, est plus satisfaisante pour l'examen des **courbes de croissance d'organismes** (notamment d'organismes supérieurs comme les vertébrés), dans ce domaine c'est un modèle concurrent du modèle de Gompertz.

Les **exemples 1 et 2** en annexe illustrent les qualités descriptives du modèle logistique pour la croissance d'un organisme (le Goëland d'Europe) et pour la croissance d'une population microbienne.

Enfin, dans la figure 3 on présente une condition complémentaire conduisant à une **solution décroissante**. Elle s'interprète comme l'action d'une substance toxique sur la biomasse, cette substance étant elle-même dégradée par cette biomasse (on peut penser, par exemple, à l'action d'un antibiotique sur une population bactérienne si celui-ci est simulanément dégradé ou métabolisé).

Le modèle exponentiel

Considérons le modèle logistique, et supposons le substrat constant (il est soit en large excédent, soit maintenu constant, grâce à une entrée dans le système, comme c'est le cas en cultures continues de bactéries). Alors on est ramené au cas simple de la croissance exponentielle.

Le modèle de Gompertz

Ce modèle a été proposé par **Gompertz** pour décrire des données actuarielles (**Gompertz**, 1825). En fait il a surtout été utilisé pour représenter la croissance de certains organismes (en particulier des vertébrés, cf. par exemple les travaux de **Laird** (1967, 1968)). Outre la bonne représentation qu'il autorise pour une variable, il permet de rendre compte du phénomène d'allométrie (cf. exemple 3 en annexe).

Dans le schéma f peut être interprété comme un facteur nécessaire à la croissance de la biomasse x (i.e. f est un facteur de croissance), comme dans le cas (S4) du modèle logistique. Ces modèles ne font intervenir que ce facteur comme élément limitant (et non un substrat).

On peut considérer, par exemple, que de telles situations sont rencontrées dans la croissance de nombreux vertébrés supérieurs : en conditions normales le "substrat" n'est pas limitant pour le jeune animal (nourriture largement fournie par les parents). On sait par contre qu'il existe un facteur hormonal de croissance: f peut représenter schématiquement un tel facteur. La loi de f est sans doute très simpliste, mais elle suffit sans doute pour approcher de nombreux exemples (ainsi on trouvera en annexe une étude de la croissance du rat musqué utilisant le modèle de Gompertz: **exemple 3**).

(ii) décroissance. Ici f peut être interprété comme un facteur de dégradation de la biomasse x, lui-même dégradé indépendamment de cette biomasse. Un exemple limite: f peut représenter un prédateur soumis à un processus exponentiel de mort, et qui consomme une proie x sans effet appréciable sur sa croissance (comparer avec le schéma du "premier cas" proposé dans le chapitre consacré aux systèmes prédateur-proie). En première approximation la dynamique de la bactérie ***Rhizobium japonicum*** a été ainsi interprétée (**Crozat**, 1983, expériences réalisées au laboratoire dans des échantillons de sol, cet exemple est présenté en annexe: **exemple 4**).

Le modèle de Kostitzin

Ce modèle a été proposé par **Kostitzin** (1937) sur la base des travaux de **Volterra** (1931) pour décrire:

- la croissance et la décroissance d'une population qui émet dans son milieu un facteur toxique;
- la croissance d'organismes pendant certaines phases de leur développement, en particulier le développement d'embryons (libres ou dans le corps maternel).

D'autres exemples peuvent être trouvés dans: **Chassé et al, 1977.**

Citons deux interprétations issues de la figure 3:

- une croissance sur un milieu limité en substrat (de type logistique), mais en présence simultanée d'un phénomène de dégradation (ou de mortalité),

- l'évolution d'une population x par consommation d'un substrat s, qui produit, en outre, un facteur toxique, ou dégradant, f. J'ai appliqué ce modèle au niveau cellulaire pour l'analyse des variations des quantités d'ARN total dans un système spécialisé: la glande sericigène du ver à soie (cf. **exemple 5**) s représentant des nucléotides, x le RNA total et f la RNAase enzyme dégradant le RNA.

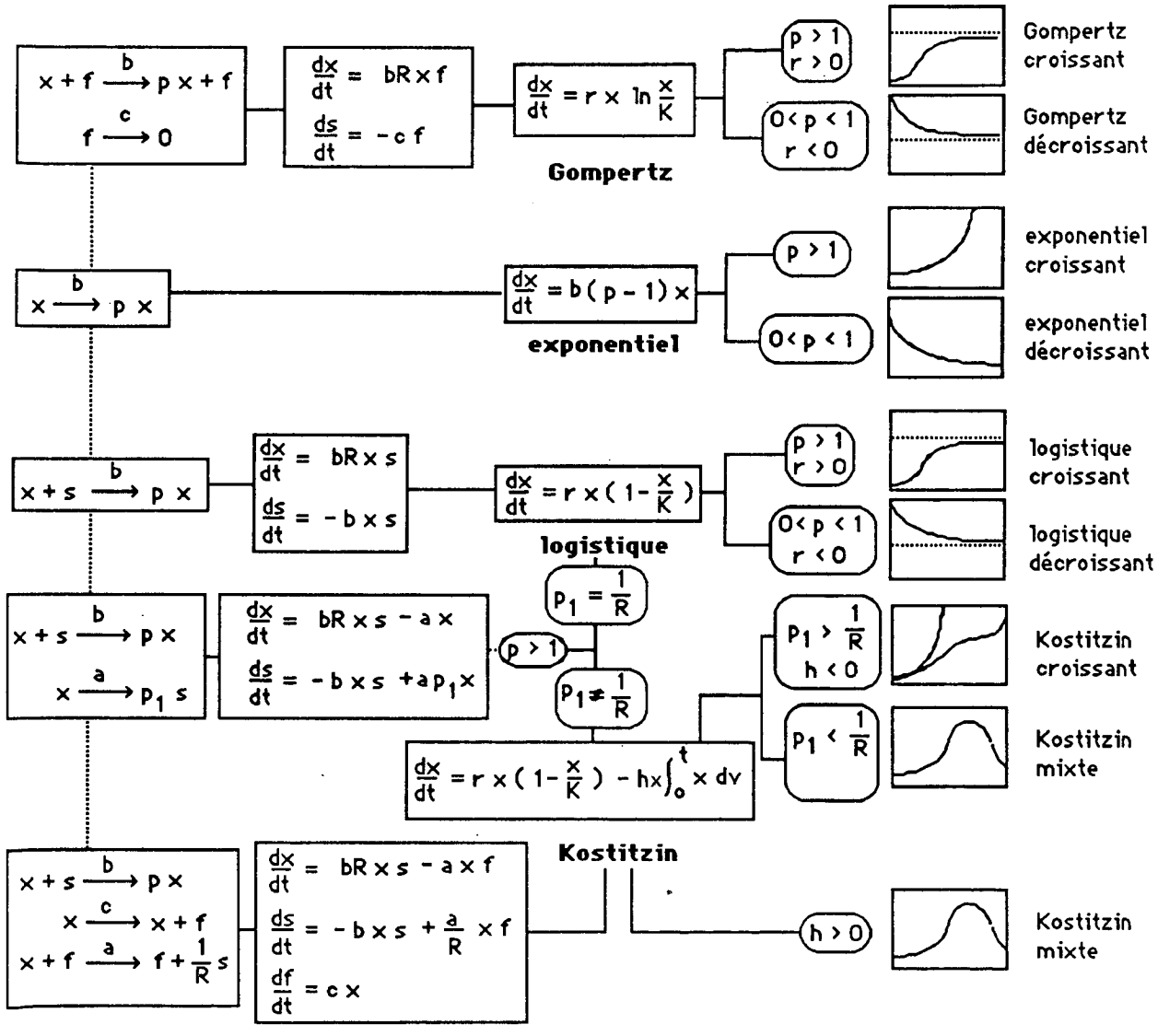


Figure 3 - Modèles simples de la dynamique des populations et leurs relations avec des schémas fonctionnels. Dans ce schéma récapitulatif tous les cas ne sont évidemment pas répertoriés, par exemple seules deux situations relatives au modèle logistique ont été reprises, nous avons retenus ceux qui nous semblent les plus intéressants dans l'optique dynamique des populations.

3.2. Modèles à deux populations en interaction

Les travaux de **Lotka** (1925, 1956) et **Volterra** (1931) sont bien connus, ils ont proposé de représenter la dynamique de populations en interaction par des systèmes d'équations différentielles. Le cas à deux dimensions correspondant à deux populations a été très étudié. De nombreux travaux mathématiques leur ont été consacrés (pour une revue générale on pourra se reporter aux ouvrages de **Pielou** (1969), de **Keyfitz** (1968), de **Lebreton et Millier** (1982), et enfin de **Oliveira-Pinto** et **Conolly** (1982)). Par contre les approches expérimentales, confrontant données et modèles, ont été plus rares; les travaux de **Gause** (1935) sont encore une référence essentielle. Il n'est pas question d'examiner toute les situations, mais plutôt de montrer comment la recherche de schémas fonctionnels peut aider à l'interprétation de modèles classiques, à leur amélioration, et même à la construction de nouveaux modèles. Ainsi on traitera, comme exemple, certains problèmes relatifs aux systèmes prédateur-proie et de compétition.

Systèmes prédateur-proie

Trois cas sont présentés en figure 4. Les deux premiers sont classiques, le troisième est à notre connaissance nouveau. Il a été écrit pour rendre compte d'un système, étudié en laboratoire, de prédation de bactéries par des protozoaires.

Premier cas: il correspond au modèle élémentaire qui décrit une croissance exponentielle de la proie x (donc sur un milieu non limitant en substrat), une croissance du prédateur y aux dépens de la proie x (x joue pour y le rôle d'un substrat), et enfin un processus de mort (exponentiel) du prédateur. Ce modèle génère les solutions oscillantes entretenues souvent citées dans la littérature. Solutions qu'il est difficile d'observer même en laboratoire. En effet si on se réfère au schéma fonctionnel il faut au moins vérifier:

-que la proie soit en croissance exponentielle (par exemple, pour des microorganismes réaliser une culture en chémostat),

-que le prédateur ait un taux de mortalité suffisant par rapport au temps de génération de la proie. Un tel processus peut être assuré par l'introduction dans le procédé expérimental d'un facteur toxique pour ces prédateurs.

Gause (1935) a tenté de telles expériences, mais les techniques étaient, à l'époque, trop rudimentaires pour permettre des observations pendant un temps assez long. Plus récemment **Bazin et Saunders** (1978) ont réalisé un "écosystème artificiel" (suivant leur terminologie) sur de telles bases, mais il n'ont pas observé les oscillations attendues (ils ont expliqué ce fait par la présence d'une singularité sous la forme d'une catastrophe du type fronce).

Deuxième cas : on suppose à présent que la proie a une croissance logistique. Ce système peut générer des solutions oscillantes amorties. Comme dans les cas précédents pour trouver le schéma fonctionnel correspondant on fait l'hypothèse que la croissance logistique est due à la présence d'un substrat limitant s, lié linéairement à x. Les différents coefficients ont été calculés pour que la matrice D du système soit de rang 2, ce qui conduit à supposer que la prédation (2ème réaction) libère dans le milieu une quantité de substrat équivalente à celle qui a été nécessaire pour la croissance de la biomasse, hypothèse peu vraisemblable (à moins de supposer que cet équilibrage est assuré par d'autres voies non explicitées dans le modèle).

Troisième cas : on suppose maintenant que les processus de prédation d'une part, et de mortalité du prédateur d'autre part, s'accompagnent d'un relargage dans le milieu d'une certaine quantité de substrat (termes en $p_1 s$ et $p_2 s$ dans les schémas fonctionnels). Cependant ce relargage n'est pas équivalent à celui consommé pour produire x. Les paramètres p_1 et p_2 sont respectivement inférieurs à $\frac{1}{R_1}$ et à $\frac{1}{R_2}$, remarquons aussi que si p_1 et p_2 sont nuls alors on retrouve le premier cas, et que si $p_1 = \frac{1}{R_1}$ et $p_2 = 0$ on retrouve le deuxième cas, il s'agit donc d'une formulation plus générale que les deux

premières. Cette extension du modèle a permis de rendre compte qualitativement de la prédation de bactéries du sol par des amibes (cf. Steinberg et al, 1987, et exemple 6). On remarque que le point d'équilibre correspond à y = 0 (i.e. disparition du prédateur).

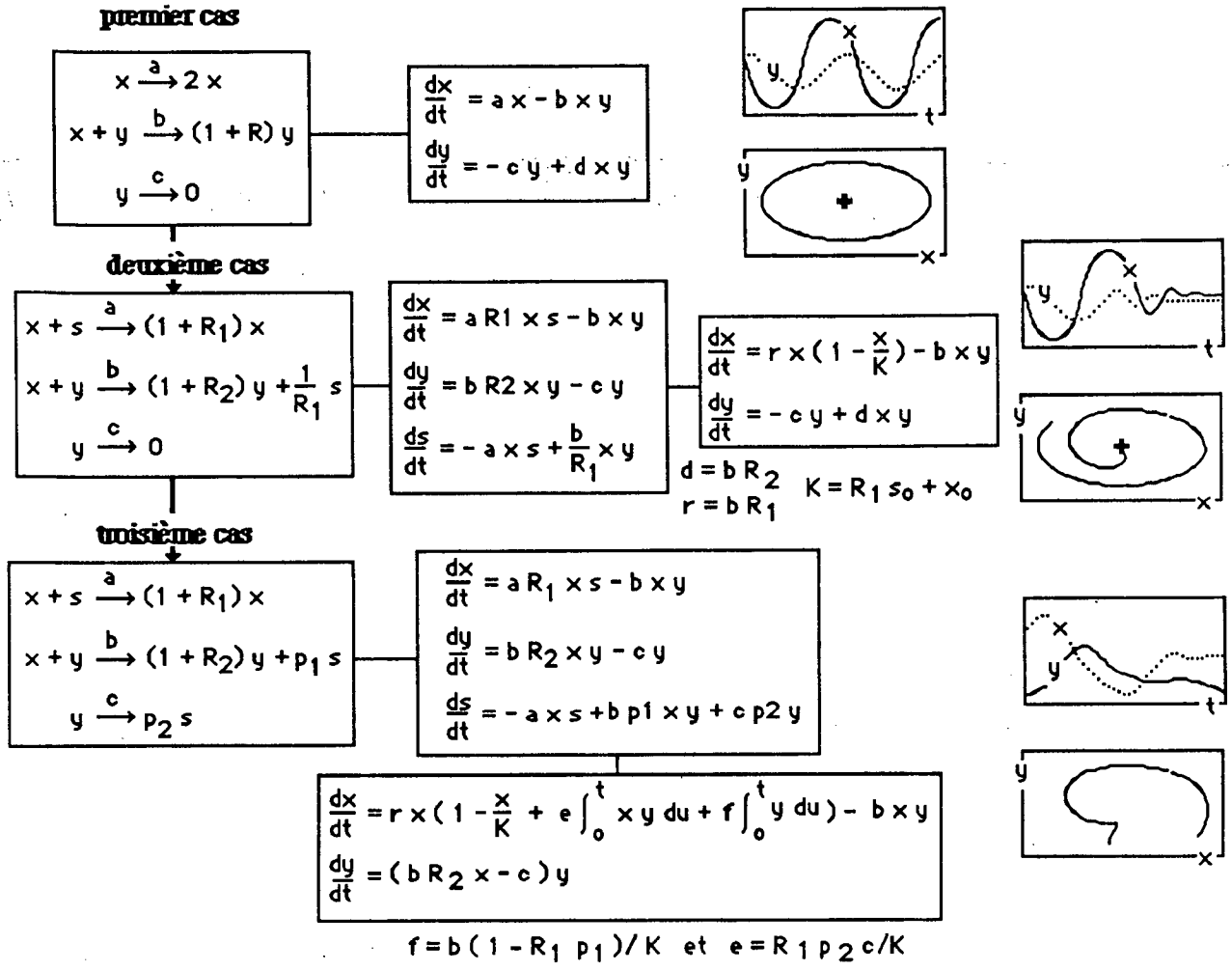


Figure 4 - Systèmes prédateurs-proie, schéma fonctionnels associés et exemples de trajectoires et de chroniques. Les équations sont en partie centrale, les schémas correspondants en partie gauche et les graphes évidemment en partie droite de la figure. Le premier cas correspond au modèle classique de Lotka-Volterra (oscillations entretenues), le second est une version qui considère non plus une croissance exponentielle mais une croissance logistique de la proie (i.e. milieu limité mais renouvelé par le relargage du substrat au cours de la prédation), le troisième suppose que le bilan du substrat (relargage - consommation) n'est plus nul (en pratique inférieur à 0). Cette dernière hypothèse a été introduite suite à l'étude d'un cas concret (système bactérie (proie) - amibe (prédateur), Steinberg et al, 1987, exemple traité en annexe).

Compétition

La compétition est généralement interprétée comme une concurrence entre des populations qui partagent des mêmes ressources (par exemple, compétition pour l'occupation de l'espace ou encore pour la nourriture ...). En ce qui nous concerne nous supposerons seulement une compétition entre deux populations x et y, pour un même substrat s sans autre interaction. Deux cas sont envisagés: le premier correspondant à des populations à mortalité négligeable (au moins pendant la durée de l'observation), le second qui fera intervenir ce processus. Classiquement ce phénomène est représenté dans la littérature par **les expressions générales:**

$$\frac{dx}{dt} = r_1 x \left(1 - \frac{x}{k_1}\right) - c_1 x y \qquad \frac{dx}{dt} = r_1 x \left(1 - \frac{x + \alpha y}{k_1}\right)$$

ou encore

$$\frac{dy}{dt} = r_2 y \left(1 - \frac{y}{k_2}\right) - c_2 x y \qquad \frac{dy}{dt} = r_2 y \left(1 - \frac{\beta x + y}{k_1}\right)$$

Dans la première formulation on reconnait des termes ressemblant au modèle logistique, et un terme d'interaction négatif censé traduire la compétition. dans la deuxième on suppose que le terme de freinage est modifié dans le modèle logistique. Les résultats sont résumés dans la figure 5, on a évidemment tenté d'introduire explicitement le partage d'une ressource commune limitante sur laquelle s'établit le phénomène de compétition.

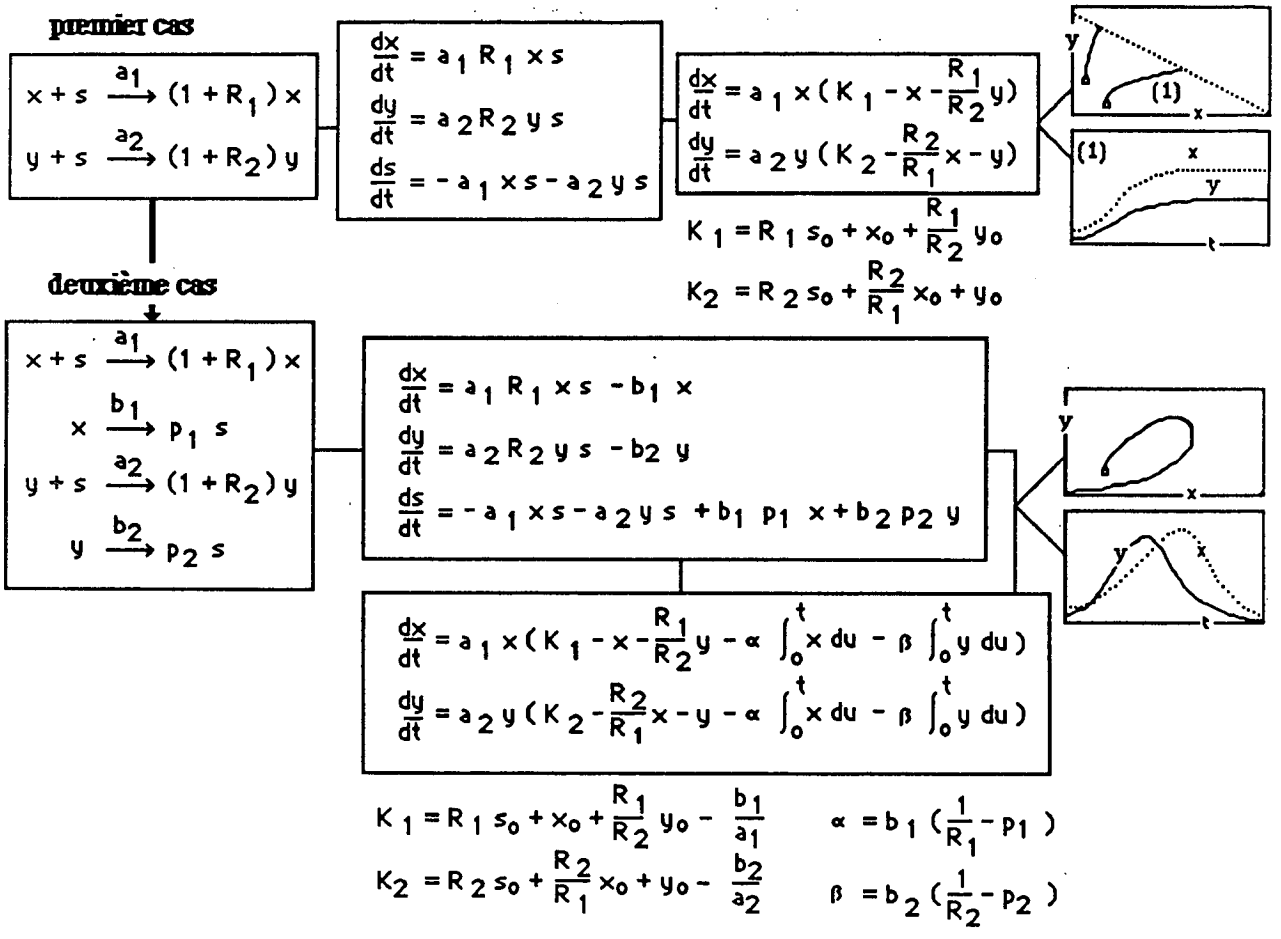


Figure 5 - Exemples de modèles de compétition de deux populations (x et y) sur une ressource limitante. Le premier cas conduit à une expression différentielle simplifiée analogue aux expressions classiques proposées ci-dessus, ce système est "dégénéré": il admet une infinité de points fixes sur la droite d'équation $y = \frac{R_2}{R_1}(K_1 - x)$, l'issue de la compétition dépend des conditions initiales. La trajectoire dans le plan (x,y) a pour équation $y = y_0 (\frac{x}{x_0})^{\mu}$ où $\mu = \frac{a_2 R_2}{a_1 R_1}$. Le deuxième cas, où une mortalité non négligeable est introduite, correspond à l'équivalent à deux dimensions du modèle de Kostitzin: les deux populations disparaissent si $p_1 < \frac{1}{R_1}$ et $p_2 < \frac{1}{R_2}$ (i.e. relargage incomplet du substrat). Dans ce dernier cas on peut trouver, suivant les valeurs des paramètres, toute une variété de solutions (par exemple, l'exclusion compétitive si le relargage est "complet": $p_1 = \frac{1}{R_1}$ et $p_2 = \frac{1}{R_2}$).

Compléments et discussion

(i) On peut envisager des procédés expérimentaux permettant de tester des représentations alternatives pour les phénomènes de compétition (par exemple pour des microorganismes). Par exemple des études de populations isolées peuvent permettre d'estimer les paramètres de la croissance (logistique, ou de Kostitzin si la mortalité n'est pas négligeable, évidemment si l'un de ces modèles convient). Ensuite une expérience de confrontation permet de tester l'hypothèse de compétition sur le substrat. C'est ainsi qu'une étude sur les problèmes de compétition a été conduite sur des souches d'un champignon microscopique pathogènes et non pathogènes pour les végétaux du genre ***Fusarium*** (**Son**, 1985). Cette étude est présentée en annexe (**exemple 7**).

(ii) Certains résultats trouvées dans la littérature peuvent se replacer dans le cadre du premier cas de compétition (cf. par exemple l'article d' **Amarger** et **Lobreau**, 1982, dans lequel des comparaisons sont faites entre les rapports des états initiaux (x_0/y_0) et ceux des états à l'équilibre (x_e/y_e) pour mettre en évidence un phénomène de compétition).

(iii) Le débat à propos du "principe d'exclusion compétitive" (hypothèse fort discutée depuis **De Bach**, 1966, et **Ayala**, 1969) peut être envisagé dans le cadre de cette approche, cependant l'étude exhaustive des différents cas de figure sortent du cadre de cet article.

(iv) Comme dans le travail de **Mac Arthur** (1970), je fais apparaître explicitement les ressources sous le forme du "substrat", mais en montrant que sous certaines hypothèses on peut en fait les rattacher à des variables implicites dans les modèles classiques de **Lotka-Volterra**.

(v) Ces modèles représentent en fait des systèmes isolés il serait intéressant de prévoir des modèles intégrant des échanges avec le milieu extérieur, c'est ainsi que l'exemple 7 est complété par une étude plus spéculative (exemple 8).

3.3 Considérations générales sur les systèmes biologiques modélisés et l'interprétation des modèles.

Hypothèses générales

Un nombre minimal d'hypothèses permet de définir un cadre restreint de travail. Les schémas fonctionnels proposés correspondent à ce cadre et n'ont aucune valeur en dehors.

(i) Les variables relatives aux **populations** (notées **x** et/ou **y**) sont les densités, ou les tailles dans un espace homogène et constant (en aire ou en volume, et relativement aux paramètres physiques du milieu), ou toutes variables proportionnelles, par exemple la biomasse ou la densité optique (d'une culture bactérienne).

(ii) Les populations sont dans un milieu isolé qui contient des **substrats limitants**. Ces substrats peuvent servir à la croissance (ressources), ce peuvent être aussi des substances toxiques (alors la biomasse est dégradée). On supposera en outre qu'un substrat est consommé. Nous nous sommes limités au cas à un substrat, mais il n'y a aucune difficulté à envisager plusieurs substrats limitants (cf. **Houllier**, dans le même volume).

(iii) A une quantité de substrat consommé correspond toujours une même quantité de biomasse produite (ou dégradée): le **rendement de la croissance** (positif ou négatif) est constant pendant la durée de l'observation.

(iv) Il peut y avoir d'autres éléments fonctionnels, qu'on appelle des **facteurs**. Ceux-ci influent sur la dynamique des populations étudiées. La différence par rapport aux substrats est que leur évolution peut être indépendante de la biomasse. En particulier ils ne sont pas consommés par la population, par contre ils peuvent être produits (c'est le cas des schémas (S4) du modèle logistique, pour n > 2).

(v) Les interactions entre variables d'état du système (i.e. variables correspondant aux populations, aux substrats, aux facteurs) sont supposées multiplicatives, c'est-à-dire relevant d'un modèle du type **loi d'action de masse (Garfinkel**, 1962), ou liées par un processus de type michaëlien (pondération de l'action de masse par un phénomène de saturation).

Interprétation: aspects phénoménologiques et mécanistes, connaissances superficielles et connaissances profondes.

Nous considérons que les modèles dans leur(s) expression(s) classique(s) représentent les **aspects phénoménologiques**: ils décrivent bien certaines situations. On peut tenter une première interprétation à ce niveau en liant les paramètres à une signification biologique. C'est par exemple le cas pour l'interprétation des paramètres r et K dans le modèle logistique. Ce niveau correspond à ce qu'on appelle **connaissance superficielle** en intelligence artificielle.

L'analyse en termes de schémas fonctionnels consiste en fait à associer à un modèle un, ou plusieurs, **processus**, c'est-à-dire tenter une explication en décomposant le phénomène en **mécanismes plus élémentaires**. En intelligence artificielle on parle alors de **connaissance profonde**.

Ainsi la présentation effectuée ici est une tentative de recherche de lien entre ces niveaux (phénoménologiques et explicatifs, connaissances superficielles et profondes). Il est néanmoins aussi clair que ce qui est phénoménologique à un niveau peut être "explicatif" ou profond pour un niveau supérieur. Ce problème d'empilage de niveaux n'est pas sans poser un problème, bien connu par ailleurs (cf. les dialogues d'Achille et la Tortue dans GEB de D. Hofsdstater, ou les tentatives de pratiques réductionnistes). On peut imaginer empiler indéfiniment des niveaux, est-ce réaliste ? Pour ma part j'aurais tendance à penser qu'on peut au plus travailler sur deux, au plus trois niveaux, tout en sachant bien que des concepts nouveaux peuvent apparaître localement. Ceci nécessite alors un bouclage entre deux niveaux, ou sur un même niveau.

Quoiqu'il en soit, pour l'instant nous proposons deux plans de référence celui des processus correspondant à une pseudo-réaction et celui des phénomènes correspondant à une ou plusieurs pseudo-réactions simultanées.

Vers une classification des modèles différentiels et intégro-différentiels de la dynamique des populations.

L'un des objectifs de ce travail était d'étudier les "liens de parenté" entre ces modèles de façon à en proposer une classification, celle-ci devant servir à la constitution de la base de connaissance d'un système "un peu intelligent" d'aide à la modélisation en biologie (projet **Edora**). L'approche par les schémas fonctionnels nous est apparu, de ce point de vue, fort intéressante (cf. également la contribution de **Houllier**, 1986, pour les modèles de croissance et développée dans ce même volume). On notera également que cette approche s'appuie sur **l'étude de situations réelles** tirées d'expériences faites au laboratoire et d'extensions raisonnables, et formalisables, pour le "terrain".

Les différentes figures proposées dans le texte montrent comment on peut construire une généalogie, précisant les liens fonctionnels entre les modèles étudiés.

Le modèle logistique apparaît comme le modèle "de base" de la croissance, on peut cependant signaler qu'une même approche pourrait parfaitement se faire en prenant le

modèle de **Monod** (1942). Il suffit de changer les règles de traduction des schémas fonctionnels (remplacer s par s/(K+s)) comme nous l'avons vu.

En tout état de cause, pour rendre ces modèles accessibles il est nécessaire d'informatiser le plus possible les manipulations formelles (et évidemment numériques). Pour ceci la description en utilisant des schémas fonctionnels apparaît bien appropriée (simplicité de la formulation, traduction automatisable...).

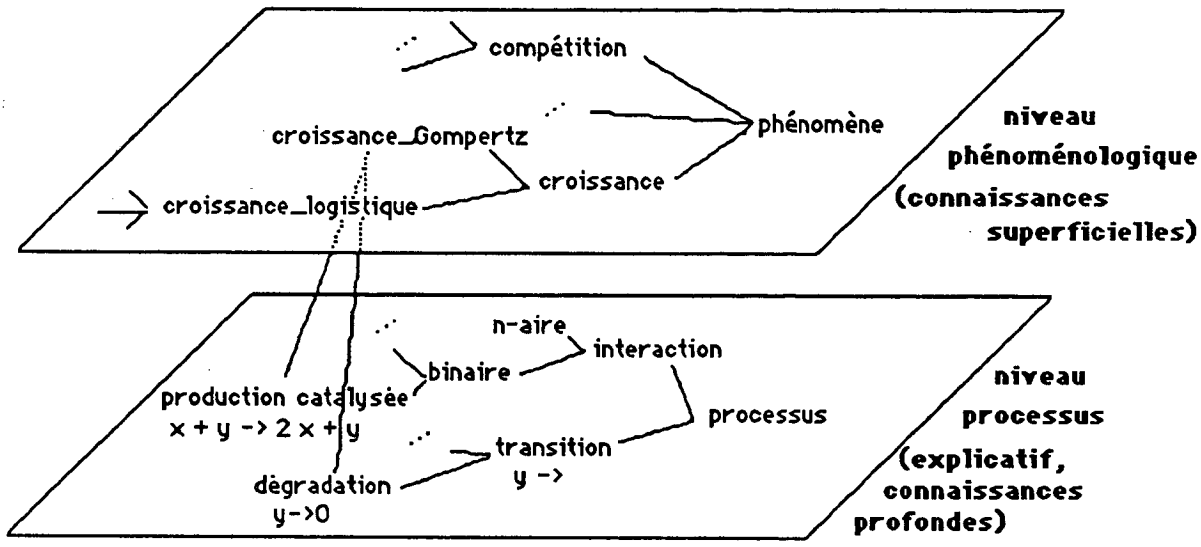


Figure 6 - Relations entre processus et phénomènes, connaissances superficielles et profondes basées sur l'analyse des modèles en termes de schémas fonctionnels. Cette figure ne présente que le principe de base des relations entre les niveaux, C. Pierret dans sa contribution fait une proposition voisine. Il y aura lieu de préciser beaucoup plus ces relations en fait plus complexe pour les inclure dans une base de connaissances opérationnelle.

4. Conclusion

Cette analyse des modèles différentiels classiques de la dynamique des populations permet de mieux comprendre leur "contenu". En particulier, un même modèle peut représenter plusieurs mécanismes au sens de schémas fonctionnels différents obtenus à partir de la traduction d'expressions algébriques équivalentes de ce modèle. La phrase de J. Monod trouve encore ici son illustration: "*le contenu d'une expression mathématique est toujours beaucoup plus riche que ne le croit en général son auteur*".

Cette approche permet aussi de préciser les relations des modèles entre eux en fonction de leur interprétation (sens fonctionnel), et également les liens avec le(s) processus élémentaire(s) qu'ils sont censés représenter, sachant qu'on peut établir d'autres relations (par exemple mathématiques) . Ces résultats et indications devraient permettre l'établissement de bases de connaissances contenant des informations, non seulement mathématiques, mais aussi phénoménologiques et explicatives.

Enfin, le formalisme utilisé permet également de

- discuter des modèles eux-mêmes et des interprétations et utilisations qui en ont été faites,

- proposer des amélioration de ces modèles qui correspondent mieux aux observations, aux résultats expérimentaux ou aux hypothèses qu'on peut émettre.

A ces fins l'utilisation de schémas fonctionnel apparaît comme efficace. Cependant je ne suis pas très satisfait du symbolisme retenu (type chimique). Je pense qu'il y aurait lieu de réfléchir à un symbolisme aussi simple mais plus parlant pour le biologiste et l'écologiste ou les symboles se rapprocheraient plus de modèles des objets et concepts qu'ils manipulent.

Bibliographie

Amarger N.S., Lobreau J.P. (1982). Quantitative study of nodulation competitiveness in *Rhizobium* strains. Appl. Environ. Microbiol., 44, 583.

Ayala F.J. (1969). Experimental invalidation of the principle of competitive exclusion. Nature, 224, 1076.

Berettra E., Vetrano F., Solimano F., Lazzari C. (1979). Some Results about Nonlinear Chemical Systems represented by Trees and Cycles. Bull. Math. Biol., 41, 641-664.

Couvreur P. (1983). Stabilité et cycles limites: le Bruxellateur. Analyse de Systèmes, 9, 2/3, 11-25.

Chassé J.L., Legay J.M., Pavé A. (1977). Le modèle de Volterra-Kostitzin en dynamique des populations. Ajustement et interprétation des paramètres. Ann. Zool. Ecol. Anim., 9, 425.

Chéruy A., Gautier C., Pavé A. (1982). Analyse des systèmes biologiques: certains aspects méthodologiques liés à la modélisation. In "La notion de système dans les sciences contemporaines" (tome 1: méthodologies). Ed. J. Lesourne, Lib. Univ. Aix.

Corman A. (1982). Modélisation du processus de nitrification dans le sol. Thèse Doct. Ing., Université Claude Bernard, Lyon 1.

Crozat Y. (1983) Caractérisation du pouvoir saprophyte des souches de *R. Japonicum* dans le sol à l'aide de l'immuno-fuorescence. Thèse Doct. Ing., Lyon.

Doolittle F.R., Hunkapiller M.W., Hood L.E., Devare S.G., Robbins K.C., Aaronson S.A., Antoniades H.N. (1983). Simian Sarcoma Virus *onc* gene, *v-sis*, is derived from the gene (or genes) encoding a platelet-derived growth factor. Sciences, 221, 275.

Emanuel N., Knorre D. (1975). Cinétique chimique. Editions MIR, Moscou, 448 p.

Hamrouni M. K. (1979). Etude et développement d'un système informatique d'aide à l'élaboration de modèles en biologie. Thèse 3ème Cycle, Université Pierre et Marie Curie, Paris.

Garfinkel D., Rutledge J. D., Higgins J. J. (1961). Simulation and Analysis of Biochemical Systems. I. Representation of Chemical Kinetics. Comm. of the ACM, 559-562, 1961.

Garfinkel D. (1962). Digital computer Simulation of an Ecological System based on a modified Mass Action Law. Ecology, 45, 502-507, 1962.

Gause G.J. (1935). Vérifications expérimentales de la théorie mathématique de la lutte pour la vie. Herman, Paris.

Gompertz B. (1825) . On the Nature of the Function Expressive of the Law of Human Mortality and a new Mode of Determining the Value of Life Contingencies. Philosoph. Transac. Roy. Soc., 115. In **Smith D. & Keyfitz N.**: Mathematical Demography. Biomath. Vol. 6, Springer-Verlag, Berlin, 1977, 279-282.

Houllier F. (1986). Construction et interprétation de modèles dynamiques: exemples forestiers. Edora 1 (même volume, p. 83-107).

Keyfitz N. (1968). Introduction to the Mathematics of Populations. Addison-Wesley, New-York.

Kostitzin V.A. (1937). Biologie mathématique. Armand Colin, Paris.

Labeyrie V. (1972). Malthusianisme et Ecologie. La Pensée, 167, 1-19.

Laird A.K. (1964). Dynamics of Tumor Growth. Brit. J. Cancer, 18, 490-502.

Laird A.K., Tyler S.A., Barton A.D. (1965). Dynamics of normal growth. Growth, 29, 249-263.

Laird A.K., Barton A.D., Tyler S.A. (1968). Growth and time: an interpretation of allometry.Growth, 32, 347.

Lebreton J.D., Millier C. (1982). modèles dynamiques déterministes en biologie. Masson, Paris.

Legay J.M. (1973). La méthode des modèles, état actuel de la méthode expérimentale. Informatique et Biosphère, Paris, 1-76.

Lotka A.J. (1925). Elements of Physical Biology. Williams & Wilkins, Baltimore.

Lotka A. J. (1956). Elements of Mathematical Biology. Dover, New-York, édition revue du précédent.

Lotka A.J. (1932). The growth of mixed populations: two species competing for a common food supply. J.Washington Acad. of Sciences, 22, 461-469.

MacArtur R. (1970). Species packing and competitive equilibrium for many species. Theor. Pop. Biol., 1, 1-11.

Monod J. (1942). Recherches sur la croissance de cultures bactériennes. Thèse Doct. ès Sciences, Herman, Paris.

Nicolis G., Prigogine I. (1977). Self-organization in Nonequilibrium Systems. Wiley, New-York.

Oliveira-Pinto F., Conolly B.W. (1982). Applicable Mathematics of Non-Physical Phenomena. Ellis Horwood & J. Wiley. Chistester.

Pavé A., Pagnotte Y. (1977). An Approach to Computer Aided Design, a Tool for Mathematical Modelling in Biology and Ecology. Comput. in Biol. and Med., 7, 301-310.

Pavé A. (1979). Dynamics of molecular populations: the evolution of RNA quantities in the silkgland during the last larval instar. Bioch., 61, 263-273.

Pavé A. (1980). Contribution à la théorie et à la pratique des modèles mathématiques pour l'analyse dynamique de systèmes biologiques. Thèse Doct. ès Sciences, Université Claude Bernard, Lyon 1.

Pavé A., Corman A. (1981). Apport de la modélisation aux études de biologie des sols: exemple de la nitrification. Sols, 4, 63-74.

Pavé A., Rechenmann F. (1985). Computer Aided Modelling in Biology: an Artificial Intelligence Approach. In "Artificial intelligence and Simulation", Soc. for Comput. Simul., Special Issue.

Pielou E.C. (1969). Introduction to Mathematical Ecology. Whiley, New-York.

Richmond B. (1985). STELLA™. User's Guide.High-Performance Systems, Dartmouth College, New Hampshire, USA, 115p.

Robertson R. (1908). On the normal rate of growth of n individuals and its biochemical significance. Archiv. fü Entwicklungs Mechan. der Organism, 25, 581-614.

Saunders P.T. (1983). Catastrophe Theory. In "Mathematics in Microbiology", Ed. M. Bazin, Acad. Press, London, 105-138.

Son M. (1985). Etude de la croissance de *Fusarium* en culture pure et en confrontation: Analyse des données expérimentales et modélisation. Rapport de D.E.A. (biologie cellulaire, biométrie). Université Claude Bernard, Lyon 1.

Steinberg C., Faurie G., Zegerman M., Pavé A. (1987). Régulation par les Protozoaires d'une population bactérienne introduite dans le sol. Modélisation mathématique de la relation prédateur-proie. *Rev. Ecol. Biol. Sol*, 24, 1, 49-62.

Verhulst P.F. (1838). Notice sur la loi que la population suit dans son accroissement. Corr. Math. Phys., 10. Trad. anglaise: A Note on the Law of population Growth. In **Smith D.& Keyfitz N.:** Mathematical Demography. Biomath. , Vol. 6, Springer-Verlag, Berlin, 1977.

Vidal C. (1978). Sur l'analyse cinétique d'un schéma réactionnel. L'Actualité Chimique, 30-72.

Volterra L. (1931). Leçons sur la théorie mathématique de la lutte pour la vie. Gauthier-Villars, Paris.

Waltmann P. (1983). Competition Models in Population Biology. CBMS-NSF Regional Conf. Series, 45, 77p.

Exemples

On trouvera dans cette annexe des exemples de données expérimentales sur lesquelles des ajustements de différents modèles exposés dans le texte ont été réalisés. La cohérence "théorique" a posteriori de chaque situation suivant les schémas fonctionnels sous-jacents est précisée. Par ailleurs, l'estimation des paramètres a été faite en utilisant la procédure VA07AD d' HARWELL, les valeurs initiales nécessaires au lancement de cette procédure on été obtenues par intégration numérique de l'équation différentielle du modèle (cf "estimation des paramètres d'équations différentielles" dans le même volume).

I - Modèle logistique

Exemple 1: croissance de *Larus argentatus*

Larus argentatus est le goëland commun d'Europe. Cet exemple s'intègre dans une étude sur la croissance des vertébrés, plus particulièrement pour mettre en évidence d'éventuelles différences entre espèces: au niveau des valeurs des paramètres de la croissance, ou même à celui du modèle; ainsi dans cette étude il apparaît nettement des croissances de type logistique, et d'autres du type Gompertz, on peut discuter de la différence sur la base des schémas fonctionnels associés. On rappelle que ce modèle s'écrit souvent:

$x' = r\,x\,(1 - \frac{x}{K})$ équation différentielle dont la solution analytique est $x = \frac{K}{1 + \frac{K-x_0}{x_0} e^{-rt}}$

Résultats

(i) données expérimentales:

Age (en jour)	0	2	5	7	10	12	15	17	20	22	25	27	30	32	35	37
masse corporelle en grammes	85	97	148	193	273	341	455	517	631	687	796	847	900	920	935	937

(ii) **estimation des paramètres:** $r = 0.1545\ j^{-1}$ $K = 992.4$ g $x_0 = 74.61$ g

(iii) **Courbe de croissance:**

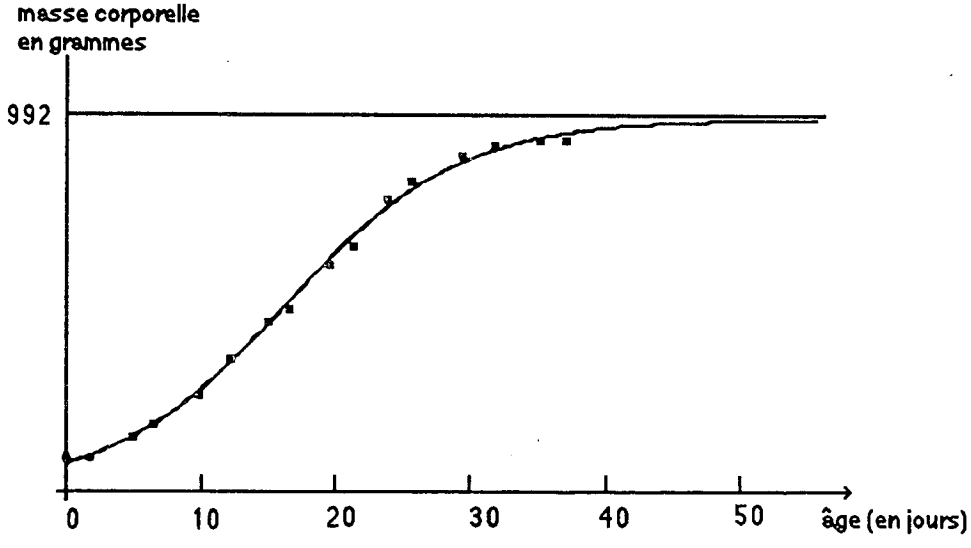


Figure e1 - Masse corporelle d'un jeune goëland en fonction de l'âge et courbe ajustée.

Conclusion

Le modèle logistique décrit bien la croissance des jeunes animaux (avant l'envol), on peut penser qu'un schéma fonctionnel du type (iii) serait satisfaisant pour représenter le mécanisme global de ce phénomène. On remarquera que le modèle de Gompertz décrit bien, lui aussi, la croissance de vertébrés supérieur . Ces deux modèles diffèrent, de ce point de vue, au niveau de l'évolution du "facteur de croissance", simple dans le cas de Gompertz, plus complexe dans le cas logistique .

Exemple 2: Croissance de populations d'Escherichia coli K 12 en milieu liquide complexe

Quantifier la croissance de populations bactériennes présente quelque intérêt, en particulier lorsqu'on désire comparer des conditions différentes de milieu. Ainsi on a utilisé une souche bactérienne bien connue, *Escherichia coli* K 12 , pour tester différents milieux de culture en laboratoire. Pour les milieux simples le modèle de croissance le mieux adapté est le modèle de Monod. Ces milieux, dits aussi minimums, sont entièrement synthétiques, ce sont les plus "simples" qui assurent une croissance, ils comportent un seul substrat limitant, ce substrat est le plus souvent un sucre. Par contre en milieux complexes, milieux comportants souvent des extraits naturels et donc plusieurs substrats utilisables par les bactéries, le modèle logistique semble suffire pour décrire la croissance de populations d'*E. coli*..

Résultats

(i) données expérimentales

Les données présentées ici proviennent d'une expérience visant à comparer divers milieux complexes couramment utilisés dans les laboratoires d'analyse médicale. Les mesures consistent à mesurer la quantité de lumière absorbée par le milieu liquide lorsque la population bactérienne croît par comparaison avec un témoin sans bactéries: l'unité de mesure est relative à la densité optique (U.D.O.), on mesure l'écart par rapport au témoin (Δ D.O.). Les conditions expérimentales choisies correspondent à une zone où la proportionnalité entre l'accroissement de la population et la variation de D.O. est vérifiée.

temps en mn	0	30	50	70	100	130	150	170	200	230	250	300	350	370	400
x en U.D.O.10^2	2.4	5.4	8.7	12.5	19.2	30.3	40.5	53.3	72.2	88.1	94.7	104.2	108.3	108.3	108.3

(ii) estimation des paramètres: $r = 0.186\ mn^{-1}$ $K = 1.10$ U.D.O. $x_0 = 0.042$ U.D.O.

(iii) courbe de croissance de la population

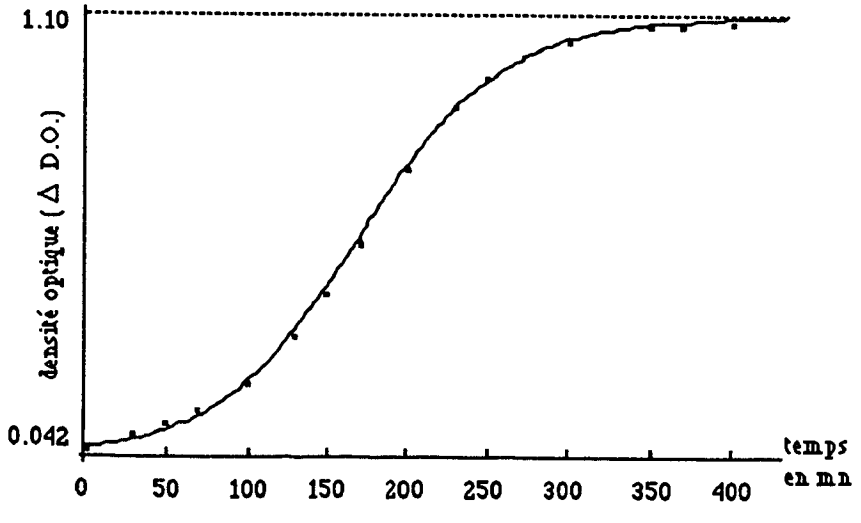


Figure e2 - Croissance d'une population d'Escherichia coli K12 en milieu liquide complexe: points expérimentaux et courbe ajustée.

Interprétation et conclusion

Le modèle logistique rend bien compte des données expérimentales, on est tenté de le voir comme représentant une croissance en milieu limité par le substrat. Cependant, pour les milieux simples (à un seul substrat limitant) le modèle de Monod est meilleur. On peut expliquer cette différence dans le modèle de réponse de la façon suivante:

D'une part, le modèle logistique est un cas limite du modèle de Monod lorsque la constante K (interprétable comme l'inverse d'une affinité des bactéries pour le substrat) est grande par rapport à la concentration initiale en substrat (ou inversement lorsque la concentration initiale en substrat est petite par rapport à cette constante): partons du modèle de Monod: $x' = \mu\ x\ s\ /\ (K+s)$ et $s' = -\ (1/R)\ x'$, si $s << K$ alors $\mu' = \mu/(K+s)$ est sensiblement constant, alors $x' \approx \mu'\ x\ s$ (cf. le modèle logistique.

D'autre part, en milieu complexe les substrats sont multiples et utilisés pour certains simultanément et pour d'autres successivement, alors on parlera plutôt de ressources au sens large. L'interprétation est plus phénoménologique: la croissance en milieu complexe limité en ressources utilisables est bien décrite par le modèle logistique. Dans les deux cas, cependant, la notion de ressources limitantes est déterminante. On prendra comme convention que la variable s apparaissant dans le schéma fonctionnel s'interprète comme une, ou un ensemble, de ressources limitantes.

II - Le modèle de Gompertz

Exemple 3 - croissance de jeunes rats musqués, *Ondatra zibethica,* (Pavé et al, 1986).

Le rat musqué colonise les régions d'étangs (par exemple la Dombes, au nord de Lyon). L'étude de son écologie conduit à s'intéresser à des caractéristiques biologiques de cet animal. Parmi celles-ci la croissance des jeunes animaux permet de préciser, par la suite, certains paramètres démographiques des populations étudiées. Ici le modèle de Gompertz apparaît mieux adapté à la représentation de la croissance de cet animal que le modèle logistique. Rappelons que ce modèle peut s'écrire:

$$x' = a \ln \frac{x}{K} \text{ et sous forme analytique } x = K \exp\left(\ln\left(\frac{xo}{K}\right)\ e^{-a\,t} \right)$$

Résultats

Des jeunes rats musqués ont été suivis depuis leur naissance (en élevage). Plusieurs variables morphologiques ont été mesurées, notamment la masse et la longueur corporelles. Ne sont retenues ici que les données relatives à un animal.

(i) données expérimentales:

Age (en jour)	masse corporelle (en gramme)	longueur de l'animal (en cm)	Age (en jour)	masse corporelle (en gramme)	longueur de l'animal (en cm)
0	16	9.1	104	688	50.3
21	116	24.6	110	695	50.5
29	175	30.3	117	712	50.7
35	264	33.4	124	739	50.9
44	352	38.6	132	728	51.0
50	416	41.3	138	747	51.2
55	447	42.4	146	733	51.3
62	503	45.4	152	738	51.3
69	540	46.5	180	763	51.7
76	540	47.7	187	757	51.8
83	603	48.3	194	765	51.9
90	646	49.3	201	767	52.0
97	684	49.9			

(ii) estimation des paramètres

La paramérisation retenue pour faire les études numériques a été: x' = a x ln(K/x), i.e., par rapport au texte a = r/ln(K), K garde la même signification (dans l'article original, on trouvera un autre exemple de paramétrisation). On a obtenu

- pour la masse corporelle: $a_m = 0.036$ et $K_m = 760$
- pour la longueur de l'animal: $a_l = 0.040$ et $K_l = 51.60$

(iii) courbes de croissance

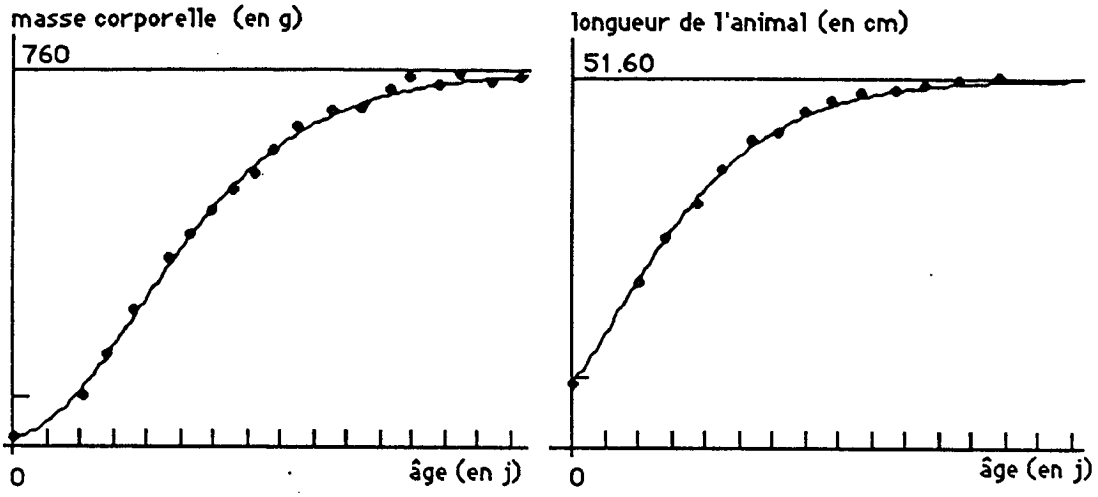


Figure e3 - Masse corporelle et longueur de l'animal en fonction de l'âge pour le jeune rat musqué. Ajustements au modèle de Gompertz et points expérimentaux (seuls quelques points ont été représentés pour rendre la figure lisible).

(iv) allométrie

Pour de nombreux organismes on observe un relation du type $y = A\ x^{\mu}$ entre des variables morphologiques, pour un même animal lors de sa croissance, ou entre animaux d'âges différents. Cette relation a été bien étudiée par G. Teissier (cf., par exemple, Teissier,1948). Si le modèle de Gompertz rend compte de la croissance, comme dans notre exemple où m représente la masse corporelle et l la longueur de l'animal,

on a $m' = a\ m \ln\left(\frac{K}{m}\right)$

supposons en outre que $l = A\ m^{\mu}$,

il vient $l' = A\ \mu\ m'\ m^{(\mu-1)}$,

soit $l' = A\ \mu\ a\ m^{\mu} \ln\left(\frac{K}{m}\right)$ et $l' = a\ l \ln\left(\frac{K^{\mu}}{m^{\mu}}\right)$

en multipliant numérateur et dénominateur du terme entre parenthèses par A alors

$$l' = a\ l \ln\left(\frac{K'}{l}\right) \text{ avec } K' = A\ K^{\mu}.$$

C'est-à-dire que ***si m suit un modèle de Gompertz de paramètre a et K et si l est liée à m par une relation d'allométrie $l = A\ m^{\mu}$ alors l suit un modèle de Gompertz de paramètre a et $K' = A\ K^{\mu}$.*** Ce résultat est illustré par la figure suivante:

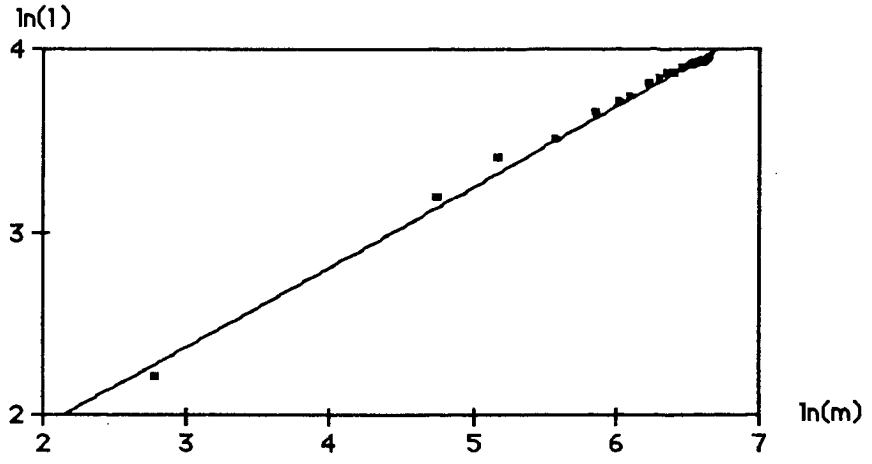


Figure e4 - Relation d'allométrie entre la longueur l du corps de l'animal et la masse corporelle m. Les valeurs des paramètres de cette relation sont: A = 2.93, μ = 0.44, K' = 52.94.

On remarquera que l'estimation de K' obtenue par régression linéaire (ln(l) sur ln(m)) est proche de celle obtenue par estimation directe (régression non linéaire de l sur l'âge).

Conclusion

La bonne qualité des ajustements obtenus montre que le modèle de Gompertz décrit bien les données expérimentales pour les variables morphologiques mesurées, la relation d'allométrie semble conservée. On peut supposer, au moins dans un premier temps, que le mécanisme de croissance suggéré par le modèle est raisonnable (i.e. processus gouverné par un facteur de croissance, celui-ci disparaissant suivant un modèle exponentiel indépendant de la biomasse produite, ce facteur est limitant, le seul dont on a besoin pour expliquer les données obtenues dans des conditions exprimentales d'élevage).

Exemple 4 - Dynamique des populations de Rhizobium japonicum dans les sols (Corman et al., 1986).

Rhizobium est un genre de bactéries qui fixe l'azote dans les sols. Ces bactéries vivent en symbiose avec ceratines plantes d'intérêt agronomique et évidemment économique (par exemple le soja). On comprend donc que l'analyse de la dynamique de ces populations bactériennes ait quelque importance. R. japonicum est une espèce exotique non présente dans les sols français à l'état naturel, c'est donc un matériel de choix pour étudier différents aspects dynamique (colonisation des sols, facteurs de régulation de ces populations...).

Résultats

L'une des premières approches expérimentales a consisté à étudier la cinétique de survie dans des sols non stériles. Il s'agissait de suivre les effectifs de populations injectées dans des échantillons de sol. Dans un premier temps plusieurs expériences ont été réalisées, elles différaient par la taille de l'inoculum initial. Les résultats semblaient pouvoir être décrit correctement par le modèle de Gompertz en remarquant que quelle que soit la valeur de l'inoculum initial le niveau de survie est le même pour un sol et une souche donnée.

(i) données expérimentales

On a retenu deux expériences correspondant à deux conditions initiales différentes:

(1) x_0=1.30 10^3

t en jours	0	2	10	20	30	45	60	75
x . 10^{-3} bact./g de sol	1.30	1.40	5.75	5.93	4.37	5.67	6.31	7.59

(2) $x_0 = 1.29\ 10^6$

t en jours	0	2	10	20	30	45	60	75
x . 10^{-3} bact./g de sol	1160	359	322	83.3	64.5	12.5	10.0	11.6

(ii) estimation des paramètres: (1) $a = 0.028\ j^{-1}$ $K \approx 10^4$ et (2) $a = 0.041\ j^{-1}$ $K \approx 10^4$

(iii) courbes

Pour rendre la figure lisible on a représenté les données et les courbes ajustées sous forme logarithmique (un calcul simple montre que si $z = \log_{10} x$, x suivant un modèle de Gompertz défini par l'équation différentielle $x' = a\ x\ \ln (K/x)$, alors $z' = a (\log_{10} K - z)$).

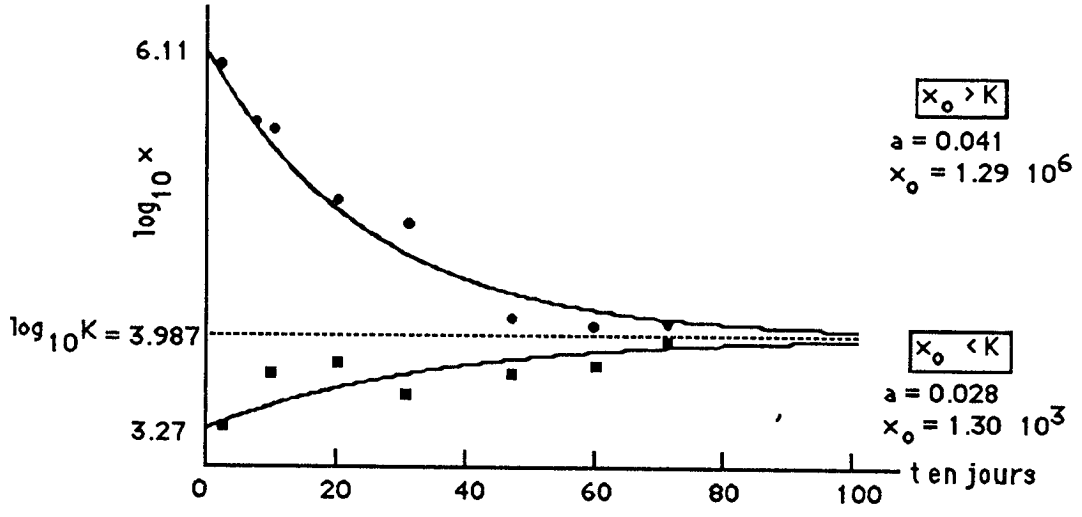


Figure e5 - Dynamique de populations de Rhizobium japonicum inoculées dans un sol non stérile. Ajustement au modèle de Gompertz $x' = a\ x\ (\ln(K) - \ln(x))$ ou $z' = a\ (\log_{10}(K) - z)$ où x est le nombre de bactéries par gramme de sol et K = 9700, les points expérimentaux sont représentés.

Conclusion

Le modèle de Gompertz décrit bien la dynamique des populations de R. japonicum dans des échantillons de sols non stériles. Remarquons au passage que le modèle logistique a également été testé, la description est beaucoup moins satisfaisante. Reste à analyser l'adéquation avec le modèle de Gompertz. En effet l'explication en terme de facteur de croissance proposée dans le texte (cf. 3.2) ne peut pas être retenue ici: K semble indépendant de x_0 et la présence d'une réponse décroissante pour $x_0 > K$ ne peut pas être expliquée par ce premier mécanisme. Cette dernière observation pourrait être interprétée par une interaction des bactéries avec un facteur "toxique" y (cf. 3.2), soit schématiquement:

$$x + y \longrightarrow y$$

$$y \longrightarrow 0$$

On peut mettre ce schéma en correspondance avec celui inféré pour un système prédateur-proie (cf. 3.2), par exemple

$$x \longrightarrow 2\ x$$

$$x + y \longrightarrow (1 + R)\ y$$

$$y \longrightarrow 0$$

où le premier processus (croissance de x, la proie) serait absent, et où le gain de croissance du prédateur y (2ème processus) serait négligeable. C'est cette observation, combinée à celle de la quasi invariance de K pour un sol donné en fonction de x_0 (et de y_0 qui peut être considérée comme constant entre expériences) qui nous a conduit à émettre l'hypothèse de régulation des populations de Rhizobium par un prédateur (Crozat, 1983, Steinberg et al, 1986): le modèle de Gompertz est alors vu comme une approximation du modèle prédateur proie sur la base d'une "proximité" des schémas fonctionnels. En fait ce raisonnement

laisse de côté le phénomène de croissance observé pour $x_0 < K$. Bien que l'hypothèse de prédation ait été vérifiée par la suite et un modèle plus précis proposé (cf. l'exemple 6, la suite de cette aventure comme exemple de système prédateur-proie), il apparaît, au moins intellectuellement satisfaisant, de reprendre le modèle de Gompertz et d'essayer de proposer un seul schéma fonctionnel permettant de représenter les deux types d'observations expérimentales. En fait, il suffit de rendre K indépendant des conditions initiales.

Reprenons l'une des formes du modèle de Gompertz:

$$x' = b\,x\,(\ln K - \ln x)$$

et supposons toujours que

$$y' = x' / (R\,x),$$

alors on a

$$y - y_0 = R\,(\ln x - \ln x_0),$$

c'est-à-dire

$$\ln x = (y - y_0)/R + \ln x_0$$

le modèle de Gompertz s'écrit alors

$$x' = b\,x\,(\ln K + y_0/R - \ln x_0 - y/R)$$
$$y' = b\,(R \ln K + y_0 - R \ln x_0 - y)$$

en posant

$$C = R \ln K + y_0 - R \ln x_0 \quad \text{et} \quad a = bC$$

le système devient

$$x' = (a/R)\,x - (b/R)\,x\,y$$
$$y' = a - b\,y$$

remarquons que $y = C$ est un point fixe du système, correspondant à $x = K$, ce dernier système différentiel admet le schéma fonctionnel suivant

$$x \xrightarrow{a/R} > 2\,x \qquad (1)$$
$$x + y \xrightarrow{b/R} > y \qquad (2)$$
$$1 \xrightarrow{a} > y \qquad (2)$$
$$y \xrightarrow{b} > 0 \qquad (3)$$

proche du modèle prédateur-proie simple, à part (2) où la croissance du prédateur est découplée du processus de prédation. Ce schéma est évidemment simpliste notamment lorsqu'on considère les rapports peu réalistes entre les constantes de vitesse, mais montre, encore une fois, la richesse d'une formulation mathématique.

III - Modèle de Kostitzin

Exemple 5: évolution de la quantité d'ARN total dans la glande séricigène du ver à soie (d'après **Prudhomme**, 1977, et **Pavé**, 1979).

A la fin du dernier âge larvaire le ver à soie, **Bombyx mori**, doit tisser son cocon de soie. Cette soie est synthétisée par la glande séricigène, notamment l'une des composantes, la fibroïne, est produite en grande quantité par les cellules de la partie postérieure de cette glande. Ce besoin de production se traduit par une grande spécialisation de ces cellules et par un fonctionnement particulier de leur arsenal protéosynthétique. Les ARN sont des molécules importantes de cet arsenal, l'étude de leurs cinétiques permet d'approcher ce fonctionnement. Outre l'intérêt évident de mieux comprendre un système économiquement important (cf. le marché de la soie), cette situation est un excellent **modèle biologique** permettant d'étudier la **différentiation cellulaire**.

On rappelle que ce modèle s'écrit: $x' = a\,x - b\,x^2 - cx \int_0^t x\,d\tau$ on peut aussi l'écrire sous forme d'un système de deux équations différentielles en posant $y = \int_0^t x\,d\tau$.

Résultats

(i) données expérimentales

temps (en jours) après la dernière mue larvaire	0	1	2	3	4	5	6	7	8	9	10	11	12
quantité d'ARN total par glande (en mg)	0.2	0.3	0.6	1.3	2.3	3.8	5.5	7.4	9.2	8.7	7.8	6.4	5.3

(ii) estimation des paramètres du modèle:

$r = 0.665\ j^{-1}$, $K = 0.0213$ mg , $c = 0.0213\ mg^{-2}$

(iii) représentation graphique

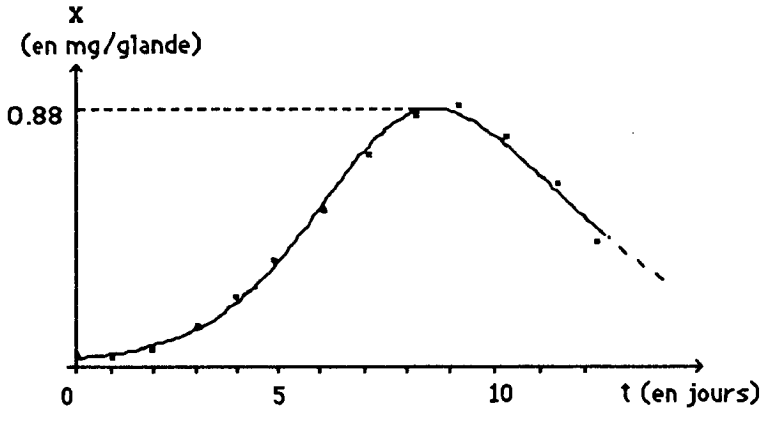


Figure e6 - Evolution de la quantité d'ARN total dans la glande séricigène du ver à soie Bombyx mori au cours du dernier âge larvaire. Courbe ajustée et points expétimentaux.

Interprétation et conclusion

On peut reprendre le deuxième schéma fonctionnel proposé pour le modèle de Kostitzin

$$x + s \xrightarrow{b} (1+R)x$$

$$x \xrightarrow{c_1} x + f$$

$$x + f \xrightarrow{c_2} f + \frac{1}{R}s$$

la première réaction représente la synthèse de l'ARN (x est la quantité d'ARN et s celle de nucléotides). La seconde réaction correspond à la production d'un facteur de dégradation f, en l'occurence une enzyme spécifique: la RNAase. La troisième réaction décrit l'action de la RNAase qui interagit avec l'ARN pour le dégrader, avec production de nucléotides (dans les mêmes proportions que pour la synthèse, hypothèse raisonnable pour ce système).

Malgré cette vision simpliste de mécanismes qu'on sait compliqués, ce modèle, sous cette interprétation, s'est révélé d'une utilisation fructueuse pour l'analyse des données expérimentales (cf **Pavé**, 1979 **op. cit** et **Pavé** 1980).

IV - Modèle de Prédation

Exemple 6: Prédation de Rhizobium japonicum par des amibes dans les sols (Steinberg et al, 1986).

Ce travail est dû à C. Steinberg, M. Zegerman s'est attaquée au problème délicat de l'estimation des paramètres du modèle. Après avoir émis l'hypothèse d'une régulation des populations de Rhizobiums par un prédateur (cf. l'exemple 4), il restait à confirmer expérimentalement cette hypothèse et parallèlement à

réfléchir sur un modèle mathématique représentant cette prédation. Une première série d'expériences confirmait cette hypothèse, notamment le suivi des populations d'amibes simultanément à celui d'une population de Rhizobiums inoculée dans un échantillon de sol. Le modèle choisi est le suivant (troisième modèle de relation prédateur-proie qui suppose une croissance limitée de la proie par la quantité de ressources disponibles et un relargage de ces ressources dans le milieu lors des processus de prédation et de mortalité du prédateur):

Schéma fonctionnel	Système différentiel
$x + s \xrightarrow{a} (1 + R_1)\, x$	$x' = a R_1 x s - b x y$
$y + x \xrightarrow{b} (1 + R_2)\, y + p_1 s$	$y' = b R_2 x y - c y$
$y \xrightarrow{c} p_2 s$	$s' = - a x s + b p_1 x y + c p_2 y$

Résultats

On se limitera à la présentation graphique des résultats d'une expérience, la représentation la mieux adaptée consiste à choisir le plan dit "de phase", c'est à dire le plan (x,y) des variables d'état. Ici les variables d'état représentent les effectifs des populations de R. japonicum (x) et d'amibes (y).

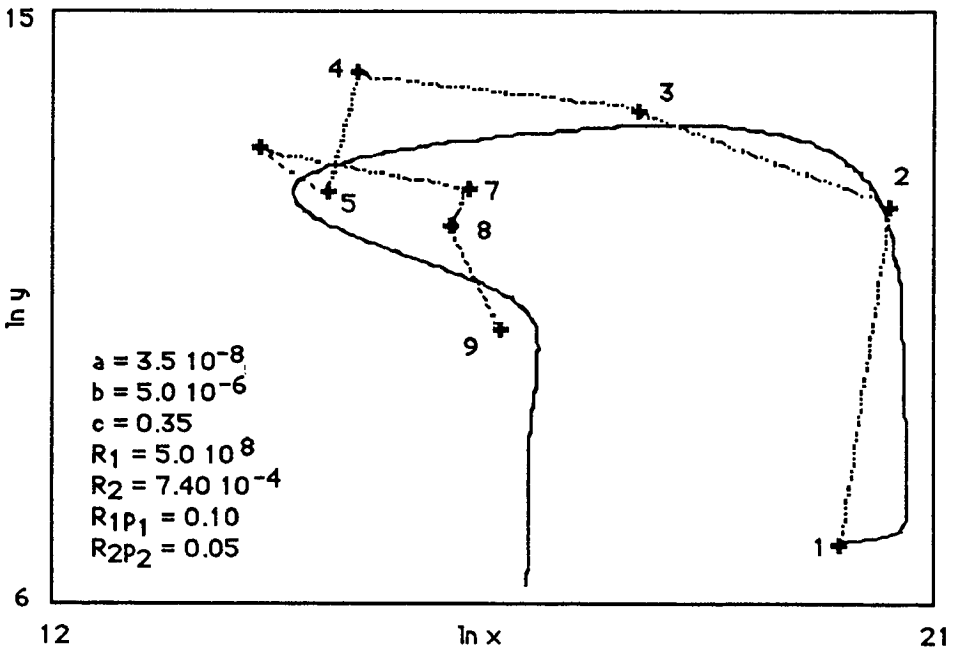


Figure e6 - Prédation d'une population de R. japonicum (x) par des amibes dans le sol. Représentation dans le plan (ln(x), ln(y)), pour des raisons de lisibilité, des points expérimentaux et d'une trajectoire obtenue par simulation.

V - Modèle de Compétition

Exemple 7: compétition de souches de Fusariums dans le sol.

Ce travail est dû à M. Zegerman, sur la base des développements théoriques présentés dans le texte (partie 3.2), elle a analysé des données proposées par R. Alabouvette et Y. Couteaudier (INRA-Dijon). Le matériel biologique est un champignon microscopique du sol, du genre *Fusarium*. Certaines espèces de ce genre sont pathogènes de végétaux, ils peuvent déclencher des maladies redoutées (les fusarioses) dans les cultures maraîchères, alors que d'autres espèces ou souches ne sont pas pathogènes.

Divers moyens de lutte existent, cependant depuis quelques temps on espère limiter les populations de pathogènes par l'apport exogène de non pathogènes en les mettant ainsi en compétition. L'objectif du

travail consiste à analyser le type de compétition, par exemple détecter un phénomène d'exclusion, autant faire se peut du pathogène. Cependant comme on va le voir dans un premier temps, la conclusion la plus probable des expériences et de l'analyse des résultats est qu'on est en présence d'un phénomène de cohabitation..., ressemblant à celui décrit dans la partie 2.2. Première conclusion qu'on pourra remettre en cause par l'analyse des conditions expérimentales que le modèle est censé représenter (système isolé): une amélioration de la formulation (système ouvert), un peu plus proche des conditions de terrain, laissant alors un peu d'espoir.

Résultats

Les résultats, résumés dans le tableau et la figure suivants, concernent deux souches (*F. oxysporum*, pathogène, et *F. solani*, non pathogène) mis en confrontation dans des échantillons de sols, au laboratoire.

	expérience 1		expérience 2		expérience 3	
temps	FS2	Foln	FS2	Foln	FS2	Foln
0	0.63	0.07	0.58	0.76	0.06	0.8
3	1.24	0.7	11.18	0.76	3.59	10.1
4	23.8	1.54	22.8	14.2	4.9	19.4
5	19.5	1.31	14.6	13.1	4.08	30
6	37.3	5.8	23.1	30.4	12.9	66
7	50	5.4	26.2	37.5	15.1	66.8
8	36.9	5.8	31.9	37.3	13.9	70
10	48.4	7.5	30.1	42	14.8	74.5
13	47.9	6.7	34.4	40	19.5	97.1
17	47.2	5.51	35.5	41.4	22.8	98
21	59.9	6.8	50.9	53.2	21.5	98
24	71	7.9	44.6	49.2	24.7	115

Le temps est exprimé en jours, les effectifs des populations en 10^4 propagules ("individus" susceptibles de se reproduire). Les expériences ont été réalisées à la station INRA de Dijon.

Conclusion

Au moins pendant la durée de l'expérience (24 jours), on peut considérer que le schéma proposé (dit de compétition pure) peut être accepté. Il s'interprète comme un partage des ressources qui conduit à une cohabitation dont le niveau ne dépend que des conditions initiales. Rappelons que ce schéma s'écrit:

$$x + s \xrightarrow{a_1} (1 + R_1)\, x \qquad \text{(x: FS2)}$$

$$y + s \xrightarrow{a_2} (1 + R_2)\, y \qquad \text{(y: Foln3)}$$

On obtient le modèle

$$\frac{dx}{dt} = a_1\, x\, (K - x - Ry)$$

$$\frac{dy}{dt} = \frac{a_2}{R}\, y\, (K - x - Ry)$$

en posant $R = \frac{R_1}{R_2}$ et avec $\mu = \frac{a_1 R_1}{a_2 R_2}$. Les deux populations croissent en se partageant le substrat disponible s. On suppose que la mortalité est négligeable pendant la durée de l'expérience.

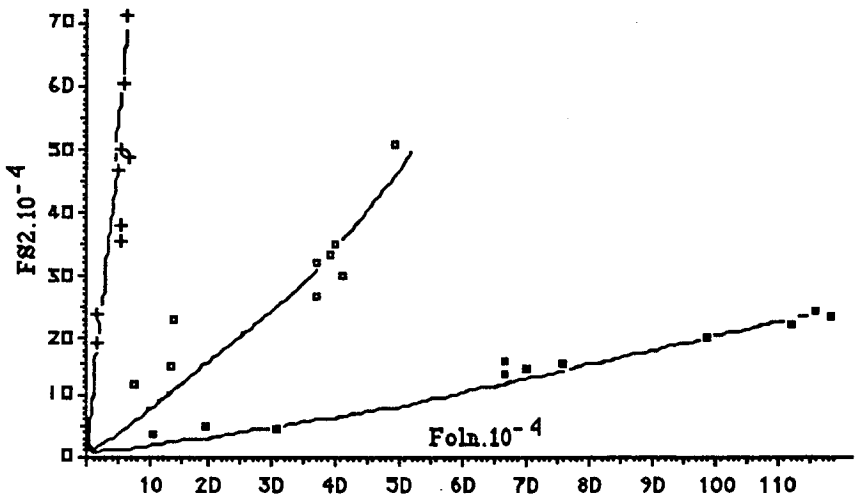


Figure e8 Compétition entre souches de Fusariums (dénomées FS2, souche de F. solanii, et Foln3, souche de F. oxysporum). Les expériences ont été menées avec les conditions initiales différentes exprimées par les rapport des tailles initiales des inoculums: $\frac{FS2(0)}{Foln(0)}$ (carrés pleins: 1/10, carrés vides: 1/1 et +: 10/1). On peut comparer les points expérimentaux et les trajectoires simulées du modèle écrit sous la forme (x: FS2 et y: Foln):

$$\frac{dx}{dt} = a_1 x (K - x - Ry)$$

$$\frac{dy}{dt} = \frac{a_2}{R} y (K - x - Ry)$$

avec $a_1 = 0.007$, $a_2 = 0.016$, $R = 1.95$ et $K = 135$. Ces valeurs ont été obtenues à partir de cinétiques en cultures pures (i.e. ne comportant qu'une seule souche à la fois) et par estimation du paramètre μ à partir de la relation $\frac{y}{y_0} = (\frac{x}{x_0})^{\mu}$ sur les résultats des différentes confrontations.

Exemple 8: étude théorique de la compétition en système ouvert.

Les cas traités jusqu'à présent, notamment dans cet exemple, concernent surtout des système isolés. Les ressources (substrat) sont consommées par la biomasse, elles peuvent éventuellement être aussi produites par celle-ci ("exploitation du milieu") mais aucun apport extérieur n'est pris en compte. Alors, s'il n'y a pas exploitation suffisante pour restaurer les ressources consommées et si les processus de mortalité ne sont pas négligeables, les populations concernées ne peuvent que disparaitre. En cas de mortalité nulle, elles cohabitent (cas dit de "compétition pure"). Ces modèles peuvent représenter des situations expérimentales comme celle étudiée ci-dessus. Par contre comment se comporte le système s'il y a une arrivée de substrat (système ouvert) ?

(i) Reprenons d'abord le schéma fonctionnel du modèle logistique auquel on ajoute un terme de mortalité (on obtient alors un schéma fonctionnel du modèle de Kostitzin):

$x + s \xrightarrow{a} (1+R)\, x$; croissance de x par consommation du substrat s

$x \xrightarrow{b} 0$; mortalité de x

Ce modèle correspond à un système isolé pour x et s , supposons maintenant que le milieu est alimenté par un flux constant de substrat, le schéma fonctionnel du système devient:

$x + s \xrightarrow{a} (1+R)\, x$; croissance de x par consommation du substrat s,

$x \xrightarrow{b} 0$; mortalité de x ,

$1 \xrightarrow{u} s$; entrée d'un flux continu constant de substrat.

Auquel correspond le système différentiel

$$\frac{ds}{dt} = - a\,x\,s + u$$
$$\frac{dx}{dt} = (a\,R\,s - b)\,x$$

On trouve facilement que le point d'équilibre est

$$s^* = \frac{b}{aR} \; ; \; x^* = u\,\frac{R}{b}$$

les isoclines (x' = 0 et s' = 0) sont:

$$s = \frac{u}{ax} \text{ et } s = s^*$$

La matrice du système linéarisé est

$$\begin{pmatrix} - a\,x^* & - a\,s^* \\ a\,R\,x^* & 0 \end{pmatrix}$$

On vérifie aisément que

si $u = 4\,\frac{b^2}{a\,R}$

alors cette matrice admet une valeurs propres réelle double, le point d'équilibre est un noeud stable,

si $u > 4\,\frac{b^2}{a\,R}$

alors cette matrice admet deux valeurs propres réelles, le point d'équilibre est un noeud stable,

si $u < 4\,\frac{b^2}{a\,R}$

alors cette matrice admet deux valeurs propres complexes, le point d'équilibre est un foyer stable,

enfin si u = 0
on retrouve évidemment le modèle de Kostitzin, si en plus b = 0 on retrouve le modèle logistique.

(ii) compétition entre deux espèces dans un système ouvert, le schéma fonctionnel peut s'écrire:

$x + s \xrightarrow{a_1} (1+R_1)\,x$	; croissance de x par consommation du substrat s
$x \xrightarrow{b_1} 0$	; mortalité de x
$y + s \xrightarrow{a_2} (1+R_2)\,y$	; croissance de y par consommation du substrat s
$y \xrightarrow{b_2} 0$	; mortalité de y
$1 \xrightarrow{u} s$	; flux constant de substrat entrant dans le milieu

on en déduit le système différentiel suivant

$$\frac{ds}{dt} = - a_1\,x\,s - a_2\,y\,s + u$$
$$\frac{dx}{dt} = (a_1\,R_1\,s - b_1)\,x$$
$$\frac{dy}{dt} = (a_2\,R_2\,s - b_2)\,y$$

La figure e9 illustre sur un exemple traité avec le logiciel DYNAMAC les différents cas de figure.

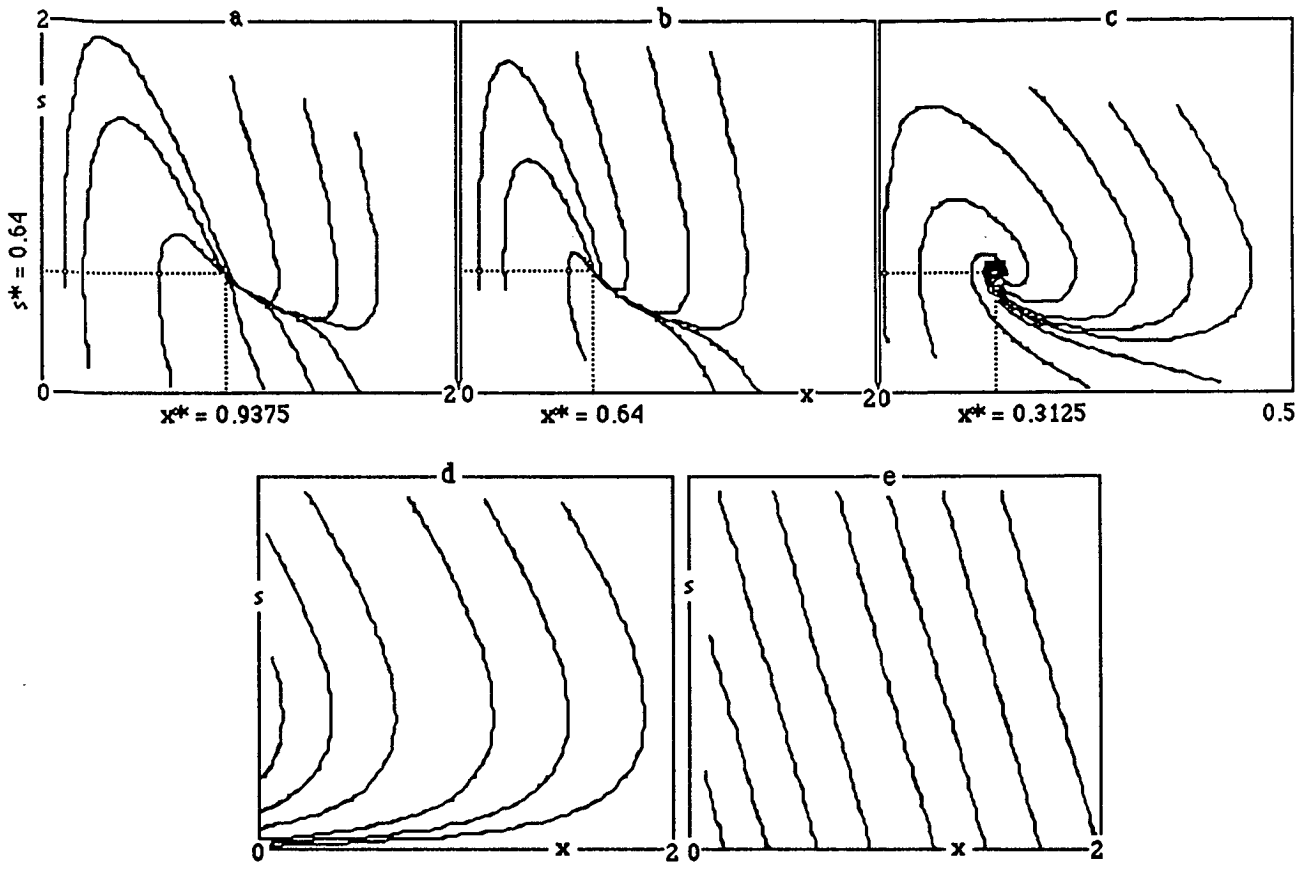


Figure e9 Modèle d'évolution d'une population dans un milieu ouvert où entre un flux constant de substrat

a et b: le point d'équilibre est un noeud stable, c: le point d'équilibre est un foyer stable

d: cas où u = 0, on retrouve le modèle de Kostitzin,

e: cas où u et b sont nuls on retrouve le modèle logistique

Ces portraits de phase ont été obtenus avec DYNAMAC [valeurs des paramètres: a = 5, R = 0.25, b = 0.8 (cas a,b,c,d) ou 0 (cas e), u = 3 (cas a) ou 2.048 (cas b) ou 1 (cas c) ou 0 (cas d et e)].

On obtient (cf. figure e10) les trois cas de figure principaux, il apparaît, à ce niveau que les hypothèses du modèle conduisent au célèbre cas de l'exclusion compétitive. On notera que la coexistence n'est possible que si $\frac{b_1}{a_1 R_1} = \frac{b_2}{a_2 R_2}$, condition très restrictive.

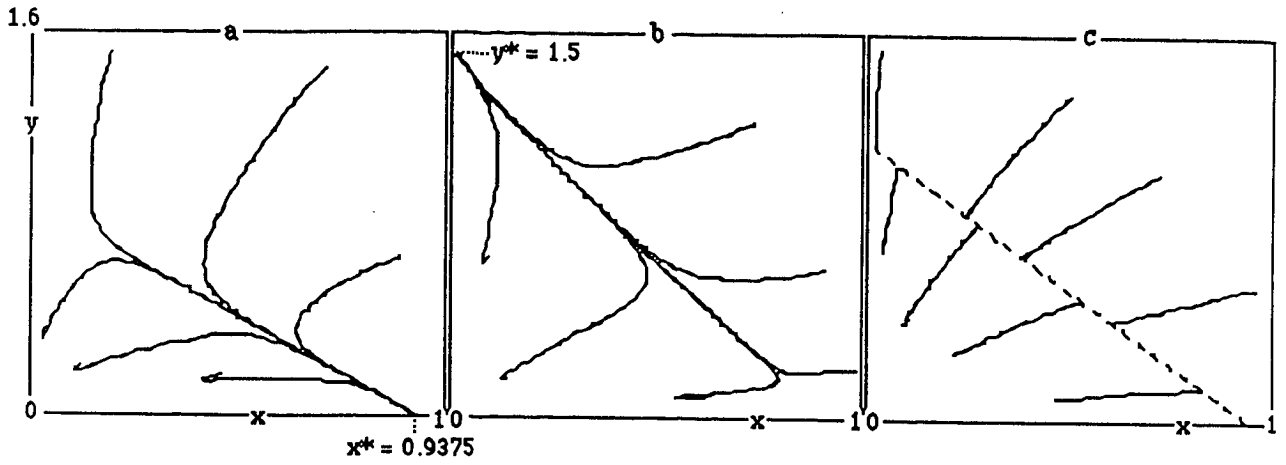


Figure e10 - Compétition de deux populations sur un même milieu, les ressources (substrat s) sont supposées renouvellées avec un flux constant u, x et y sont des quantités proportionnelles à la densité des populations considérées. On observe en a et b le phénomène d'exclusion compétitive, en c la coexistence entre les populations. Ces figures sont obtenues en jouant sut le taux de mortalité b_2 de y. Dans le cas a on a $b_2 = 0.7$, pour b on a $b_2 = 0.4$, pour c on a $b_2 = 0.512$ qui correspond à la valeur telle que $x^* = y^* = \frac{b_1}{a_1 R_1} = \frac{b_2}{a_2 R_2}$. Par ailleurs, les paramètres correspondant à x sont les mêmes que ci-dessus, pour y on a $a_2=4$, $R_2=0.2$, enfin $s_0=6$ et $u=3$.

Ce modèle rend compte de ce qui peut se passer en milieu ouvert, dans un cas encore simple mais sans doute plus proche des conditions naturelles que les modèles de systèmes isolés. En particulier, si on se réfère à l'exemple précédent, il reste possible d'observer le phénomène d'exclusion compétitive en milieu naturel pour le fusarium. Cet exemple de la compétition montre
- d'une part qu'il faut bien préciser à quelles conditions une représentation mathématique représente un système biologique,
- d'autre part que le schéma fonctionnel permet de mieux formaliser les hypothèses, et donc de mieux saisir ces conditions.

Ainsi, il faut éviter de tirer des conclusions hâtives ou trop générales à partir d'un modèle mal explicité, mal interprété, ou représentant une situation spécifique et dont on cherche à tirer des conclusions hors du cadre strict des hypothèses de sa construction. Par exemple, pour la compétition, il aurait été prématuré de conclure à l'absence du phénomène d'exclusion compétitive, en milieu naturel, à partir de modèles correspondants à des systèmes expérimentaux isolés. Ceux-ci conduisent à prévoir une coexistence (mortalité négligeable pendant la durée de l'expérience) ou une disparition des deux populations (mortalité non négligeable). Par contre, le modèle pour un système ouvert simple proposé est sans doute un peu plus proche de conditions naturelles, il semble indiquer que le phénomène d'exclusion compétitive pourrait jouer un rôle important. Il serait d'ailleurs intéressant de simuler expérimentalement un tel système. Par ailleurs, on pourrait compliquer un peu le modèle en testant un flux de substrat non constant dans le temps, i.e. u = f (t) en supposant seulement que f(t) est bornée et positive, éventuellement on pourra évacuer cette dernière hypothèse et supposer qu'outre un apport il puisse y avoir un flux sortant.

CONSTRUCTION ET INTERPRETATION DE MODELES DYNAMIQUES: Exemples forestiers [1]

François HOULLIER
Laboratoire de Biométrie, Université Claude Bernard - Lyon 1
et Station de Sylviculture et de Production, C.R.F. Champenoux, I.N.R.A.

La représentation mathématique de l'évolution des peuplements forestiers fait appel à de nombreux types de modèles dynamiques. Je m'intéresse ici à deux groupes de modèles qui permettent d'illustrer les différents types de démarche au moment de la ***conception*** d'un modèle.

J'examine en premier lieu le cas des ***courbes de croissance***. Le formalisme de ***la cinétique chimique*** est utilisé, soit pour ***construire*** de nouveaux modèles à partir de la description discursive du processus de croissance, soit pour proposer des ***interprétations*** *a posteriori* de modèles classiques.

Les limites de ce formalisme sont analysées après un bref rappel du principe général et je propose un certain nombre de pistes pour son extension. Ce formalisme permet par ailleurs de déterminer ***des relations de proximité entre modèles***, basées sur la parenté des représentations fonctionnelles pseudo-chimiques associées.

J'examine en second lieu le cas d'un ***modèle démographique global*** pour une population d'arbres groupés en stades qui relève d'une approche plus empirique. Deux modèles en temps discret, l'un déterministe, l'autre stochastique, sont associés aux ***représentations graphiques*** des flux d'arbres d'un stade à un autre et de la séquence des stades possibles pour un même individu.

Les problèmes liés à l'***estimation*** des paramètres de ces modèles sont ensuite discutés à la lumière de ces représentations.

1. Introduction

La représentation mathématique de l'évolution des peuplements forestiers [2] fait communément appel à de nombreux types de modèles dynamiques (**Munro**, 1974; **Dudek** et **Ek**, 1980). Dans la plupart des cas, l'évolution des peuplements est décomposée en plusieurs processus, chacun de ces processus étant représenté par un modèle: courbe de croissance, modèle de mortalité, scénario de prélèvement. Mais il existe parallèlement des modèles globaux qui ne procèdent pas à une telle décomposition.

---

[1] Une partie de ce texte avait été initialement conçue sous forme de questions posées aux écophysiologistes et avait été présentée au "sous-groupe Modélisation" du "Groupe Sylviculture" du Département Forêts de l'I.N.R.A.; l'objectif était de rechercher auprès des biologistes des connaissances susceptibles d'orienter ou de légitimer le choix d'un modèle de croissance en dehors des seuls critères statistiques.

[2] Le lecteur non averti de la terminologie trouvera en annexe une définition concise de ces termes.

Le problème du choix d'un modèle se pose donc, dans ce domaine comme dans d'autres, à plusieurs niveaux: choix de la forme globale du modèle (décomposition en sous-modèles ou modèle unique), de sa nature (probabiliste ou déterministe, en temps discret ou continu), choix entre différentes fonctions,... (voir par exemple **Chéruy, Gautier et Pavé**, 1982). On distingue alors classiquement deux grands types d'approches (**Legay**, 1973; **Ek** et **Dudek**,1980):

- ***L'approche "empirique" (situation --> modèle --> théorie)*** : il s'agit de chercher, dans un ensemble de modèles pris a priori (exemples: différents monomes en vue d'une régression linéaire multiple ou une collection de courbes de croissance classiques), celui qui produit le meilleur ajustement, au sens statistique, à un jeu particulier de données expérimentales; la qualité de l'ajustement est par exemple mesurée par la somme des carrés des écarts entre observations et prédictions. Le modèle obtenu est généralement bien adapté à la situation particulière étudiée. Cette approche correspond souvent, et paradoxalement, à des objectifs de prédiction; quelques difficultés inhérentes à cette pratique sont en effet:
. l'absence fréquente de signification physique ou biologique des paramètres du modèle: on s'intéresse plus aux relations structurelles qu'aux mécanismes fonctionnels;
. la possibilité de toujours améliorer l'ajustement en ajoutant des paramètres ou des variables "explicatives";
. la sensibilité souvent élevée du modèle par rapport à ses paramètres et la difficulté d'extrapoler l'ajustement hors de la zone couverte par les données expérimentales.

- ***L'approche "conceptuelle" (théorie --> modèle --> situation)***, aussi appelée approche "théorique": il s'agit idéalement de déduire la forme du modèle à partir de considérations théoriques relatives aux propriétés physiques ou biologiques du phénomène étudié et aux propriétés mathématiques des modèles possibles. Un des problèmes rencontrés "en forêt", et dans d'autres domaines, est que, si l'on veut tenir compte de toute la connaissance (souvent qualitative) accumulée, on est rapidement confronté à un nombre excessif de facteurs et on est amené à faire:
. soit des impasses biologiques (si les mécanismes d'action des facteurs sont mal connus);
. soit des impasses mathématiques (multiplication du nombre de paramètres, "instabilité du modèle").

L'approche conceptuelle est souvent utilisée quand le modèle sert comme outil d'investigation de la connaissance. Mais elle peut aussi s'avérer utile pour "contraindre" des modèles à but prédictif à avoir un comportement "raisonnable" en dehors du domaine défini par les données disponibles (**Ek** et **Dudek**, 1980).

Ces deux approches ne sont en fait pas exclusives, et il est aussi important de choisir un modèle proche de la réalité expérimentale et qui ne soit pas qu'un objet abstrait que de chercher un modèle cohérent avec des connaissances antérieures ou extérieures aux seules données utilisées pour l'ajustement ou l'identification. On peut enfin envisager une troisième approche: celle du "mathématicien", qui ne considère pas un objet biologique particulier, mais qui s'intéresse aux modèles du point de vue de leurs propriétés mathématiques afin, par exemple, de suggérer des interprétations possibles dont la validité devra ensuite être jugée dans un contexte particulier (exemple: "la théorie des catastrophes").

Le lien entre le phénomène biologique représenté et son modèle formel, en particulier sa représentation mathématique, peut se faire grâce à divers types de représentations (**Pavé**, 1980). Selon l'approche adoptée, ces représentations, soit servent directement à la construction du modèle, soit permettent d'interpréter un modèle déjà construit. Ces différentes démarches sont illustrées ici, dans le cadre forestier et pour deux types de modèles:

- ***Courbes de croissance*** d'individus ou de populations: on cherche, au moyen d'un ***formalisme de type chimique***: (1) à (ré)interpréter un certain nombre de modèles différentiels classiques (**Debouche**, 1979; **Lebreton** et **Millier**, 1982), fréquemment utilisés dans le domaine forestier, en termes de mécanismes possibles, (2) à générer des modèles voisins en recombinant des mécanismes élémentaires dont l'existence est vraisemblable.

- ***Modèle démographique global*** pour des populations structurées en stades: ce modèle n'est pas basé sur des mécanismes à proprement parler biologiques, mais sur la représentation graphique de la séquence des états possibles pour un arbre ou sur le bilan des "flux d'arbres" groupés en classes de diamètre à l'intérieur d'une population.

2. Construction et interprétation de courbes de croissance

2.1 - Courbes de croissance et formalisme de la cinétique chimique

La biologie de la croissance des individus ou des populations met en oeuvre un ***ensemble complexe de mécanismes*** (consommation de nutriments, vieillissement, interaction, coopération et compétition, prédation). On se place ici dans le cadre de la dendrométrie classique: c'est-à-dire qu'on cherche à représenter, à des fins pratiques, (gestion forestière) la dynamique d'ensemble d'un arbre ou d'un peuplement. Dans ce cadre, un modèle (représentant par exemple la croissance en hauteur d'un arbre) ne peut pas prendre en compte tous les mécanismes biologiques sous-jacents; d'une part, parce qu'ils ne sont pas tous identifiés ou quantifiés; d'autre part, parce que la complexité mathématique d'un tel modèle le rendrait inutilisable (besoin d'un grand nombre de mesures, inadaptation aux objectifs), voire "dangereux" (parce qu'instable, par exemple).

On est donc amené à rechercher des modèles plus simples qui intègrent simultanément plusieurs processus élémentaires. On souhaite cependant savoir si ces modèles ont une certaine signification biologique. Cette possibilité d'interprétation permet en effet d'analyser qualitativement les capacités prédictives de ces modèles en dehors des données pour lesquelles ils ont été conçus, de formuler des hypothèses, voire de susciter des expériences afin de tester ces hypothèses.

L'introduction du formalisme de la cinétique chimique en biologie est principalement due à **Garfinkel** et **al** (1962) et **Garfinkel** (1968). Des applications antérieures existaient cependant déjà avec en particulier le modèle enzymatique de Michaëlis (voir par exemple **Millier**, 1982). Ce formalisme a ensuite été complété par un algorithme de passage réciproque d'un système d'équations différentielles multilinéaires à un système de réactions chimiques, puis il a été appliqué à des populations de molécules ou pour réinterpréter certains modèles en temps continu de la dynamique des populations (**Pavé** et **Pagnotte**, 1977; **Pavé**, 1980; **Pavé** et **Rechenmann**, 1986).

Ce formalisme n'est, bien sûr, qu'une représentation simplifiée de la réalité; décrire une croissance par des mécanismes pseudo-chimiques ne signifie pas que ces mécanismes ont bien lieu, mais que ***nos connaissances ou nos observations actuelles sont compatibles avec un tel résumé***. Trois exemples simples permettent d'illustrer cette remarque:

. le formalisme n'interdit pas de faire "réagir" simultanément plusieurs "variables"; on sait pourtant en chimie que des réactions impliquant plusieurs molécules sont hautement improbables;

. à la place des réactifs chimiques, on utilise des variables qui peuvent aussi bien décrire l'effectif d'une population, que les caractéristiques morphologiques d'un individu, des quantités de matière (nutriments, eau) ou des flux d'énergie;
. dans le domaine forestier, la plupart des processus biologiques suivent des cycles complexes et imbriqués (journaliers et annuels par exemple): en dendrométrie classique, on néglige les cycles intraannuels et on considère la croissance comme un phénomène "régulier".

Le formalisme chimique s'apparente donc à un "simulateur" ou à un "modèle vrai" (**Legay**, 1973) selon que les réactions pseudo-chimiques du modèle représentent ou non la logique de la croissance, c'est-à-dire selon la validité de l'analogie entre d'une part le schéma fonctionnel (les réactifs et les réactions) et, d'autre part, les composantes du système et les interactions de ces composantes. Les mécanismes obtenus doivent en fait être considérés comme des ***mécanismes globaux*** résultant de la superposition de mécanismes élémentaires d'importance variable: l'existence d'un facteur limitant du milieu peut ainsi déterminer l'essentiel de la forme du modèle et occulter largement d'autres mécanismes ou d'autres facteurs (cf. § 1.6).

2.2 - Cas général

On se donne un système de p réactions chimiques élémentaires mettant en présence q réactifs:

$$n_{i1} X_1 + n_{i2} X_2 + \dots + n_{iq} X_q \xrightarrow{a_i} m_{i1} X_1 + m_{i2} X_2 + \dots + m_{iq} X_q \quad (1 \le i \le p);$$

on lui associe (**Emanuel** et **Knorre**, 1975, p.150, et **Pavé** dans le même volume) le système d'équations différentielles ordinaires:

$$\frac{dX_i}{dt} = \sum_{i=1}^{r} a_i (m_{ij} - n_{ij}) \prod_{k=1}^{q} X_k^{n_{ik}}$$

avec les conditions initiales: $X_j(0)=X_{j,0} \ge 0 \quad (1 \le j \le p)$ (dans la suite ces conditions initiales seront omises).

. X_i est le $i^{ème}$ élément (réactif); par analogie, x_i désignera une variable morphométrique décrivant un individu (ex: hauteur d'un arbre), ou la quantité d'un produit (ex: nutriment), ou encore l'effectif d'une population;
. on peut aussi représenter une "source" (ex: immigration) ou un "puits" (ex: mortalité, émigration) extérieurs par une constante non nulle (cf. § 2.3.3 et 2.3.4);
. n_{ij} et m_{ij} sont les coefficients stoechiométriques de la réaction: en cinétique chimique ce sont des entiers, mais ici on les suppose réels (cf. § 2.4.3);
. a_i est la constante de vitesse de la ième réaction;
. $m_{ij}-n_{ij}$ est le bilan pour l'élément j de la réaction i.

Le formalisme pseudo-chimique est particulièrement bien adapté à la recherche de ***modèles déterministes différentiels***. Dans certains cas, il est cependant possible d'intégrer le système différentiel et d'obtenir un modèle analytique sous la forme $X = f(t)$ ou $t = f^{-1}(X)$.

2.3 - Exemples élémentaires

On considère maintenant quelques exemples simples qui illustrent comment ce formalisme permet de construire, à partir de réactions "simples", des modèles plus complexes (cf. § 1.6 et 1.7). On désignera généralement par X la variable qui décrit la croissance de l'individu ou de la population étudiés et par S ou F les facteurs qui favorisent ou inhibent cette croissance.

2.3.1 - *Croissance par consommation d'un substrat*

La réaction "chimique": $X + S \xrightarrow{a} (1+R)\,X$ décrit la croissance d'un individu ou d'une population, caractérisés par X (ex: biomasse), sur un substrat S (ex: élément minéral du sol); cette réaction a un rendement R et une constante de vitesse a. Le système différentiel associé est:

$$\frac{dX}{dt} = a \,.\, R \,.\, X \,.\, S$$
$$\frac{dS}{dt} = -\,a \,.\, X \,.\, S \qquad \text{(cf. § 2.5.1).}$$

Dans la pratique, il est rare qu'on puisse limiter la description du milieu à un seul substrat. On a plutôt tendance à le décomposer en différents facteurs. Si on suppose que la croissance repose sur la consommation simultanée de deux facteurs (la généralisation au cas de plusieurs facteurs est immédiate, voir § 1.6), on peut proposer la "réaction":

$X + S_1 + S_2 \xrightarrow{a} (1+R)\,X$ auquel est associé:

$$\frac{dX}{dt} = a \,.\, R \,.\, X \,.\, S_1 \,.\, S_2$$
$$\frac{dS_1}{dt} = \frac{dS_2}{dt} = -\,a \,.\, X \,.\, S_1 \,.\, S_2$$

2.3.2 - *Dégradation spontanée et mortalité*

Dans un certain nombre de cas, le substrat utilisé disparaît ou se dégrade, apparemment, de lui-même (ex: lessivage des éléments minéraux dans un sol), ou bien on observe des phénomènes de mortalité naturelle. Pour décrire ce type de phénomènes, on utilise la pseudo-réaction: $X \xrightarrow{a} 0$, dont l'équation différentielle associée est : $\frac{dX}{dt} = -\,a.X$, où a est la vitesse relative (taux) de dégradation de X. Si X représente la biomasse d'un arbre ou d'un peuplement forestier, la réaction peut décrire une chute de litière ou la mort d'un certain nombre d'arbres.

2.3.3 - *Apport extérieur d'un substrat*

Le cas opposé se produit quand il existe une source d'énergie ou de matière extérieure au système étudié. Des exemples immédiats sont, en biologie végétale, la lumière et l'eau et, en biologie animale, les phénomènes migratoires pour des populations mobiles. Pour représenter de tels apports, on peut proposer la réaction $1 \xrightarrow{a} X$, dont l'équation différentielle associée est : $\frac{dX}{dt} = a$, où a est le flux de l'apport extérieur.

2.4 - Compléments sur le formalisme pseudo-chimique

Les quelques exemples précèdents montrent la souplesse de ce formalisme et sa capacité à décrire des situations diverses. Cependant, avant d'aller plus loin dans la construction et l'interprétation de courbes de croissance, il est nécessaire de préciser certaines de ses propriétés. Une présentation complète peut être trouvée dans **Pavé** (1980): formulation matricielle, algorithme d'obtention d'un schéma chimique à partir du système différentiel (c'est-à-dire résolution du problème d'interprétation pseudo-chimique d'un modèle différentiel). On préfère donc illustrer ici certains points sensibles liés à l'utilisation de ce formalisme.

2.4.1 - *Disymétrie des membres gauche et droit et réversibilité des réactions*

Le signe '+' n'a pas le même sens dans les deux membres des réactions et ne doit pas être assimilé à l'addition classique. Par exemple, les deux schémas suivants sont équivalents du point de vue du bilan, mais les systèmes différentiels associés sont très différents:

$$X + 2S \xrightarrow{a} (1+R)X + S \qquad X + S \xrightarrow{b} (1+R)X$$

$$\frac{dX}{dt} = a \cdot R \cdot X \cdot S^2 \qquad dX,dt) = b \cdot R \cdot X \cdot S$$

$$\frac{dS}{dt} = -a \cdot X \cdot S^2 \qquad \frac{dS}{dt} = -b \cdot X \cdot S$$

Il est aussi possible d'utiliser des réactions réversibles; cela revient à écrire deux réactions de sens contraire ayant chacune une constante de vitesse propre; par exemple:

$$X + S \underset{a}{\overset{b}{\dashleftarrow\!\!\dashrightarrow}} Y \qquad \frac{dX}{dt} = \frac{dS}{dt} = -\frac{dY}{dt} = -a \cdot X \cdot S + b \cdot Y$$

Ces deux exemples montrent que l'***écriture d'un bilan n'est, en toute rigueur, pas suffisante***: il faut en effet aller plus loin et décrire, au moins partiellement, les ***mécanismes de croissance***.

2.4.2 - *Non unicité de l'interprétation réciproque et identi-fiabilité des modèles*

Si on se donne un système d'équations différentielles du type de celui du § 1.2, on peut trouver au moins une interprétation sous forme de schéma fonctionnel. Mais cette interprétation n'est pas nécessairement unique (voir **Pavé**, même volume et le modèle de Kostitzin). Il faut donc être prudent quand on essaie d'interpréter un système différentiel sous forme de schéma fonctionnel, puis de mécanismes biologiques.

Cette pluralité d'interprétations possibles doit être rapprochée des problèmes d'identifiabilité des modèles. En effet, dans la pratique, on ne dispose souvent que:
. d'une série chronologique portant sur la variable X;
. de connaissances qualitatives sur les facteurs du milieu et leurs mécanismes d'action.

On ne peut donc pas estimer tous les paramètres mis en jeu par le schéma fonctionnel et ***on ne juge les performances du modèle que sur la base du comportement de la seule variable X***. Plusieurs schémas voisins peuvent donc souvent être proposés sur la base de ces seules connaissances.

Exemple: on dispose d'une série $(X(t_i), i=1, \ldots, N)$ et on considère le schéma du § 1.3.1; on obtient, par intégration de $\frac{dX}{dt} = - R \frac{dS}{dt}$:

$$\frac{dX}{dt} = a.X.(M-X) \quad \text{où} \quad M = X(0) + R.S(0);$$

on remarque alors qu' on ne peut estimer que les paramètres a, M et X(0): R et S(0) restent inaccessibles.

2.4.3 - *Effet d'une relation d'allométrie et problèmes d'unité*

Si l'individu étudié peut être caractérisé par deux variables X et Y en relation allométrique l'une avec l'autre (**Bertalanffy**, 1973, p.168), $X = k \,.\, Y^{r}$ (par exemple, X et Y peuvent être la biomasse et le diamètre ou la hauteur d'un arbre), alors on voit sur l'exemple du § 1.3.1 que:

$$\frac{dY}{dt} = a \,.\, \frac{R}{r} \,.\, Y \,.\, S$$
$$\frac{dS}{dt} = - (a \,.\, k) \,.\, Y^{r} \,.\, S$$

Ce système différentiel peut être généré par:

$$Y + S \xrightarrow{a} (1 + R')\, Y + S \quad \text{où} \quad R' = \frac{R}{r}$$
$$r\,Y + S \xrightarrow{a'} r\,Y \quad \text{où} \quad a' = a \,.\, k$$

Le choix de ***la variable qui caractérise l'individu*** n'est donc pas neutre pour la représentation du phénomène de croissance. Le choix des ***unités*** est, lui aussi, essentiel puisque les schémas fonctionnels suivants ne sont pas équivalents du point de vue des mécanismes de croissance et des systèmes différentiels:

$$2\,X + 2\,S \xrightarrow{a} 2\,(1 + R)\,X \qquad X + S \xrightarrow{b} (1 + R)\,X$$

$$\frac{dX}{dt} = 2 \,.\, a \,.\, R \,.\, X^2 \,.\, S^2 = - R \,.\, \frac{dS}{dt} \qquad dX,dt) = b \,.\, R \,.\, X \,.\, S = - R \,.\, \frac{dS}{dt}$$

2.4.4 - *Dépendances en fonction du temps et phénomènes aléatoires*

Dans ce qui précède, on a supposé que les schémas fonctionnels étaient invariants au cours du temps: les paramètres a et R du § 1.3.1, étaient ainsi considérés comme fixes. On peut cependant imaginer que la vitesse des réactions et leur rendement varient au fil du temps. Ce genre de variation peut par exemple traduire des phénomènes de vieillissement ou les fluctuations de facteurs non clairement identifiés ou non pris en compte dans le schéma fonctionnel (exemple: climat). L'introduction de dépendances temporelles déterministes complique sensiblement le système différentiel qui n'est plus nécessairement autonome (cf §1.7.2).

De la même façon qu'on peut imaginer des dépendances déterministes des schémas en fonction du temps, on peut considérer que les réactions ont un certain caractère aléatoire (par exemple, dû aux fluctuations climatiques) issu de facteurs non explicitement décrits dans le schéma fonctionnel. La prise en compte de phénomènes aléatoires peut être envisagée de deux façons:

- en considérant les paramètres des réactions (rendements, constantes de vitesse) comme des processus stochastiques;
- en remplaçant les équations différentielles déterministes par des équations différentielles stochastiques (**Garcia**, 1983).

2.4.5 - *Cas des retards et des seuils*

Les modèles "à retard ou à seuil" sont assez fréquemment utilisés (**Millier**, 1982). Le formalisme pseudo-chimique permet, dans une certaine mesure, de générer ce type de modèle. Pour rendre compte de délais, on peut introduire des facteurs intermédiaires. Par exemple, si S=S(t) est imposé de façon extérieure (exemple: réserve en eau d'un sol):

$$X + S \xrightarrow{a} (1 + R)\,X \quad \text{donne} \quad \frac{dX}{dt} = a \,.\, R \,.\, X \,.\, S(t)\ ;$$

pour introduire un retard, on peut dissocier cette réaction en deux réactions élémentaires:

$$S \xrightarrow{b} S' \qquad \text{ou} \qquad S + S' \xrightarrow{b} S''$$
$$X + S' \xrightarrow{b'} (1 + R')\,X \quad (1) \qquad X + S'' \xrightarrow{b'} (1+R')\,X \quad (2)$$

$$\frac{dS'}{dt} = b \,.\, S(t) - b' \,.\, X \,.\, S' \qquad \frac{dS'}{dt} = - \,b \,.\, S \,.\, S'$$
$$\frac{dX}{dt} = b' \,.\, R' \,.\, X \,.\, S' \qquad \frac{dS''}{dt} = b \,.\, S \,.\, S' - b' \,.\, X \,.\, S''$$
$$\frac{dX}{dt} = b' \,.\, R' \,.\, X \,.\, S''$$

d'où:

$$S'(t) - S'(0) = b \int_0^t S(u)\,du \; - \; \frac{1}{R'}\,[\,X(t) - X(0)] \qquad \text{pour (1)}$$

ou $$S''(t) - S''(0) = - \,[S'(t) - S'(0)] \; - \; \frac{1}{R'}\,[X(t) - X(0)] \qquad \text{pour (2)}$$

puis $$\frac{dX}{dt} = (b'.R').\ X\ .(F(0) + G(t) - \frac{X}{R'})\,, \qquad \text{avec:}$$

pour (1): $$F(0) = S'(0) + \frac{X(0)}{R'} \qquad \text{et} \qquad G(t) = b \int_0^t S(u)\,du$$

pour (2): $$F(0) = S''(0) + S'(0) + \frac{X(0)}{R'} \qquad \text{et} \qquad G(t) = S'(0).e^{\,-\,(b.\,\int_0^t S(u)\,du)}$$

Cette équation peut alors s'interpréter comme une croissance logistique (croissance par consommation d'un substrat; cf § 2.3.1 et 2.5.1) avec dépendance de l'asymptote en fonction du temps (l'asymptote valant F(0) + G(t)).

Les modèles à seuil sont moins faciles à générer (exemple: la fermeture du couvert forestier peut être considérée comme un seuil, car elle correspond à un changement d'état du système et à la saturation du "substrat lumineux" cf § 2.7.2). Une solution peut être d'insérer dans le schéma fonctionnel des variables intermédiaires générant des variations rapides de la courbe de croissance. Une autre solution est de considérer deux schémas fonctionnels distincts de part et d'autre du seuil, mais on peut douter de la validité de cette démarche qui introduit une discontinuité biologiquement discutable et qui pose de plus des problèmes d'identification (les procédures classiques requièrent en effet des critères continument dérivables par rapport aux paramètres; cf **Vila**, 1982).

2.5 - Modèles classiques

On cherche ici à utiliser le formalisme mis en place, pour proposer des interprétations de nature mécaniste sous forme de schémas fonctionnels pour quelques modèles classiques. Pour les modèles logistique et de Gompertz, on reprend les interprétations proposées par **Pavé** (1986).

2.5.1 - *Modèle logistique*

Sous forme différentielle, le modèle logistique s'écrit: $\frac{dX}{dt} = a.X.(M-X)$ où M est l'asymptote du modèle. Le schéma fonctionnel et le système différentiel associé ont été donnés au § 2.3.1. (croissance par consommation d'un substrat) et repris au § 2.4.2.

2.5.2 - *Modèles de Gompertz, de Johnson-Schumacher et de Lundqvist-Matern*

Sous forme différentielle, le modèle de Gompertz s'écrit: $\frac{dX}{dt} = A.X.\ln\frac{M}{X}$; pour en obtenir une représentation pseudo-chimique, on introduit un élément extérieur, F tel que $F(t) = A.\ln\frac{M}{X(t)}$, soit $\frac{dF}{dt} = -\frac{A}{X}\frac{dX}{dt}$ et $F(0) = A.\ln\frac{M}{X(0)}$; on a alors le système:

$\frac{dX}{dt} = X.F$ qui est associé à: $X + F \xrightarrow{a} (1+R)X + F$, avec a.R=1

$\frac{dF}{dt} = -A.F$ $F \xrightarrow{b} 0$, avec b=A.

Une interprétation possible de ce modèle est donc celle de la croissance déterminée par un facteur de croissance (hormone en biologie ou catalyseur en chimie) qui se dégrade indépendamment du processus de croissance proprement dit; un tel mécanisme peut par exemple décrire un phénomène de vieillissement. Si on ne dispose que de mesures concernant X, alors on ne peut estimer que les paramètres X(0), A et M; on ne peut donc pas estimer individuellement a et R, mais seulement leur produit.

Sous forme différentielle, le modèle de Johnson-Schumacher (**Debouche**, 1979) s'écrit: $\frac{dX}{dt} = A.X.(\ln \frac{M}{X})^2$ (avec A>0); sa forme intégrée est: X(t) = M.exp(1/[B-A.t]), pour t > B/A et avec $B = \frac{1}{\ln [\frac{X(0)}{M}]}$. Comme précédemment, on introduit F tel que: $F=\ln (\frac{M}{X})$:

$\frac{dX}{dt} = A \,.\, X \,.\, F^2$ qui est associé à: $X + 2\,F \xrightarrow{a} (1+R)\,X + 2\,F$, avec a.R=A

$\frac{dF}{dt} = -\,A \,.\, F^2$ $\qquad 2\,F \xrightarrow{b} 0$, avec 2.b=A.

L'interprétation proposée pour le modèle de Gompertz est ici aussi possible, mais on observe une différence dans les proportions "stoechiométriques" de la croissance et donc dans le détail du mécanisme de croissance. La valeur de l'asymptote est définie par: $M = X(0).\exp [\frac{a.R.F(0)}{2.b}]$.

Sous forme différentielle, le modèle de Lundqvist-Matern (**Debouche**, 1979) s'écrit: $\frac{dX}{dt} = A.X.[\ln (\frac{M}{X})]^C$ (avec A > 0 et 1 < C < ∞); sa forme intégrée est: $X(t) = M.\exp[\frac{-1}{(B-A.t)^D}]$, pour t > B/A et avec $B = \frac{1}{\text{Log}[\frac{X(0)}{M}]}$ et $D = \frac{1}{(C-1)}$. On introduit F, tel que $F(t) = \ln (\frac{M}{X(t)})$;

$\frac{dX}{dt} = A \,.\, X \,.\, F^C$ qui est associé à: $X + C\,F \xrightarrow{a} (1 + R)\,X + C\,F$, avec a.R=A

$\frac{dF}{dt} = -A \,.\, F^C$ $\qquad C\,F \xrightarrow{b} 0$, avec b.C=A.

Ce modèle généralise les deux précédents: le modèle de Johnson-Schumacher correspond à D=1, soit C=2; le modèle de Gompertz apparaît comme la "limite" (pour le schéma fonctionnel) du modèle de "Lundqvist-Matern" lorsque C tend vers 1.

2.5.3 - *Modèle monomoléculaire*

Sous forme différentielle, le modèle monomoléculaire, ou modèle de Mitscherlisch, s'écrit: $\frac{dX}{dt} = a - b.\,X$. On peut l'obtenir directement à partir du cas général:

$1 \xrightarrow{a} X$

$X \xrightarrow{b} 0$.

Ce modèle peut donc s'interpréter comme la superposition (1) d'un processus linéaire de croissance par apport extérieur d'un substrat et (2) d'un processus exponentiel de décroissance (mortalité). Sous cette forme, il n'est pas immédiat d'imaginer une interprétation forestière de ce modèle qui est par exemple employé par **Maugé** (1975) pour décrire la croissance du Pin maritime dans les Landes. On peut toutefois proposer une interprétation plus proche de mécanismes biologiques vraisemblables en introduisant un substrat intermédiaire S renouvelable:

$1 \xrightarrow{a} S$

$S \xrightarrow{b} X$ soit: $\frac{dX}{dt} = b \,.\, S - c \,.\, X$

$X \xrightarrow{c} 0$ $\frac{dS}{dt} = a - b \,.\, S$

d'où $S(t) = (S(0) - \frac{a}{b}).e^{-b.t} + \frac{a}{b}$ et $X(t) = K.e^{-c.t} + K'.e^{-b.t} + K''$.

(avec: $K' = \frac{b.S(0) - a}{c - b}$, $K'' = \frac{a}{c}$ et $K = X(0) - K' - K''$)

On retrouve alors le modèle monomoléculaire, de façon exacte dans les cas où:

. $K' = 0$, $K < 0$ et $K'' > 0$,

. $K = 0$, $K' < 0$ et $K'' > 0$,

. $b = C$, $K + K' < 0$ et $K'' > 0$,

et de façon approximative si K' est négligeable devant X(t) (condition suffisante).

2.5.4 - *Modèle "logistique généralisé"*

Le modèle logistique généralisé est parfois appelé modèle de Nelder ou modèle de Chapman-Richards. Il généralise les modèles logistique, monomoléculaire et de Gompertz (**Lebreton** et **Millier**, 1982). Sous forme différentielle, il s'écrit:
$\frac{dX}{dt} = A.X - B.X^m$, avec A et B de même signe.

Ce modèle a aussi été proposé par **Bertalanffy** (1973), à partir de considérations théoriques facilement transposables dans le formalisme pseudo-chimique; le modèle différentiel $\frac{dX}{dt} = A.X - B.X^m$ est obtenu comme la superposition d'un processus d'anabolisme et d'un processus de catabolisme qui s'expriment comme des fonctions allométriques de X (**Pienaar** et **Turnbull**, 1973). Ce modèle est par exemple associé au schéma fonctionnel:

$X \xrightarrow{a} 2\,X$ (autoreproduction)

$m\,X \xrightarrow{b} 0$ (mortalité).

Il est possible de générer ce modèle à partir d'autres schémas fonctionnels (cf § 2.5.1, 2.5.2 et 2.5.3 pour les cas particuliers des modèles logistique, de Gompertz et monomoléculaire). Si on reprend le schéma fonctionnel du modèle logistique (§ 2.3.1) et si on suppose qu'il existe une relation d'allométrie $X = k.Y^r$, on obtient un schéma équivalent (du point de vue de la croissance de X ou Y) à celui donné au § 2.4.3. et on peut écrire l'équation différentielle:

$\frac{dY}{dt} = a \,.\, \frac{k}{r} \,.\, Y \,.\, (M'^{\,r} - Y^{\,r})$, où $M' = [\frac{M}{k}]^{1/r}$ (voir aussi **Bailly**, 1985).

On considère maintenant le modèle voisin $\frac{dX}{dt} = A.X^n.(M-X)$ ($A \geq 0$ et $M \geq 0$); il est associé au schéma fonctionnel: $n\,X + S \xrightarrow{a} (n + R)\,X$, avec $M = R.S(0)+X(0)$ et $a = A$. Ce schéma est voisin de celui du § 2.5. Il en diffère par la valeur des "coefficients stoechiométriques" de X. L'ordonnée du point d'inflexion vaut $M\frac{n}{n+1}$.

2.6 - Modèles à plusieurs "substrats"

2.6.1 - *Croissance "logistique" sur plusieurs substrats*

A partir de l'exemple du § 1.3.1, on considère le cas où la croissance est déterminée par plusieurs substrats, au nombre de N, et on s'intéresse au schéma:

$$X + S_1 + S_2 + ... + S_N \xrightarrow{a} (1 + R)\,X\ ;$$

on obtient le système différentiel:

$$\frac{dX}{dt} = a\,R\,X \prod_{i=1}^{N} S_i \quad = -\,R\,.\,\frac{dS_j}{dt} \quad , 1 \leq j \leq N;$$

soit $M_i = X(0) + R\,.\,S_i(0)$, on a alors: $\frac{dX}{dt} = a\,.\,R^{(N-1)}\,.\,X\,.\prod_{i=1}^{N}(M_i - X)$.

Ce modèle est proche du modèle de Blumberg (**Lebreton** et **Millier**, 1982, p.23): $dX/dt = C.\,X^{a}.\,(M-X)^{b}$, et constitue une autre forme de généralisation du modèle logistique. On a choisi ici de supposer que les différents substrats intervenaient simultanément, mais on peut certainement rencontrer des situations où ils interviennent successivement ou de façon plus complexe.

Dans le cas où N=2, on montre que si $0 < X(0) < M_1 \leq M_2$:

- X(t) est croissant, admet une asymptote fixée par M_1 et un unique point d'inflexion entre 0 et M, d'ordonnée $X^\circ = (M_1+M_2-(M_1^2+M_2^2-M_1.M_2)^{1/2}) / 3$;
- X° peut varier entre $\frac{M_1}{3}$, quand M_2 tend vers M_1, et $\frac{M_1}{2}$, quand M_2 tend vers l'infini.

On observe ainsi que : (1) un facteur non limitant peut intervenir dans la croissance (ici, au niveau de la vitesse de croissance par le facteur (M_2-X)); (2) dans certains cas, un schéma peut être simplifié sans grande perte: quand M_2 est très grand devant M_1, alors la croissance de X suit approximativement le modèle logistique.

Dans le cas où N > 2, on montre encore que pour $0 < X(0) < \mathrm{Min}(M_j) = M$, X(t) est croissante, admet une asymptote supérieure M et un point d'inflexion unique entre 0 et M. L'ordonnée X° de ce point d'inflexion est comprise entre $M/(N+1)$ si $M_j = M$ ($1 \leq j \leq N$), et $\frac{M}{2}$ si un seul facteur est limitant et si tous les autres sont largement surabondants. Il n'existe pas d'expression analytique de X en fonction de t, mais une expression réciproque de t en fonction de X.

2.6.2 - *Facteur de croissance et substrats multiples*

La prise en compte de plusieurs substrats n'est généralement pas suffisante pour décrire certains phénomènes de croissance. Dans le domaine forestier, on sait par exemple que le vieillissement est un mécanisme essentiel. On a vu, au § 1.5.3, qu'il était possible de rendre compte de ce genre de situations en introduisant un facteur de croissance. On peut donc, par exemple, considérer:

$$X + F + S \xrightarrow{a} (1 + R)\,X + F \qquad \frac{dX}{dt} = a \,.\, R \,.\, X \,.\, S \,.\, F$$

auquel on associe:

$$\frac{dS}{dt} = - a \,.\, X \,.\, S \,.\, F$$

$$F \xrightarrow{b} 0 \qquad \frac{dF}{dt} = - b \,.\, F$$

d'où: $\frac{dX}{dt} = a \,.\, X \,.\, (M - X) \,.\, F(0) \,.\, e^{-b.t}$, avec $M = R.S(0) + X(0)$. On peut obtenir une expression analytique de t en fonction de X.

Dans le cas de plusieurs substrats, on construit de la même façon le schéma:

$$X + F + S_1 + S_2 + ... + S_N \xrightarrow{a} (1 + R)\,X + F$$
$$F \xrightarrow{b} 0$$

2.7 - Modèles avec substrat renouvelable

Un certain nombre d'éléments nécessaires à la croissance des arbres sont sujets à des flux extérieurs: il s'agit par exemple de la lumière, de l'eau ou de certains nutriments (**Harper**, 1977). Ces éléments peuvent s'accumuler dans le système et rester disponibles pour la croissance, ou être évacués quand ils ne sont pas consommés. L'interprétation proposée pour le modèle monomoléculaire est un premier exemple de ce type de situations.

2.7.1 - *Substrat renouvelable accumulable*

Le schéma:

$$X + S \xrightarrow{a} (1 + R)\,X$$
$$1 \xrightarrow{b} S$$

décrit un mécanisme de croissance sur un substrat renouvelable qui s'accumule dans le milieu. Le système différentiel associé est:

$$\frac{dX}{dt} = a \,.\, R \,.\, X \,.\, S$$

$$\frac{dS}{dt} = - a \,.\, X \,.\, S + b = - \frac{1}{R} \,.\, \frac{dX}{dt} + b$$

d'où: $\frac{dX}{dt} = a.X.(M(t) - X)$, avec $M(t) = R.S(0) + X(0) + R.b.t$. Ce modèle admet une asymptote oblique définie par $X=M(t)$ et un point d'inflexion unique entre 0 et cette asymptote oblique.

2.7.2 - *Substrat renouvelable non accumulable*

Pour décrire un substrat non accumulable, on peut compléter le schéma précédent par: $S \xrightarrow{c(t)} > 0$, où c(t) est tel que $\frac{dS}{dt} = - a.X.S + b - c(t).S = 0$.
Cette équation s'apparente donc à une contrainte sur la quantité du substrat S disponible à chaque instant pour la croissance de X: on a S(t) = S(0) pout tout $t \geq 0$. En considérant que la production de X, mesurée par $\frac{dX}{dt}$, et la consommation instantanée de S pour la croissance, ΔS, sont liées par $\Delta S = - \frac{1}{R} \frac{dX}{dt}$, on obtient alors:

. $\frac{dX}{dt} = a.R.S(0) . X$, si $b-a.X.S(0) > 0$: domaine où le substrat n'est pas limitant;

. $\frac{dX}{dt} = R.b$, si $b-a.X.S(0) < 0$: domaine où le substrat est limitant.

Ce modèle présente un ***seuil*** défini par $t_S = \frac{1}{a.R.S(0)} Ln[X(1)-X(0)]$ et $X_S = \frac{b}{a.S(0)}$: en-deçà, la croissance est exponentielle; au-delà, elle est linéaire. Si on a de plus $Y = k.X^r$ (relation d'allométrie), on observe, pour Y:

. une première phase de croissance exponentielle $Y(t) = Y(0).e^{a.R.r.S(0).t}$, pour $t < t_S$;

. suivie d'une seconde phase d'allure parabolique $Y(t) = k . (X_S + R.b.[t-t_S])^r$, pour $t \geq t_S$.

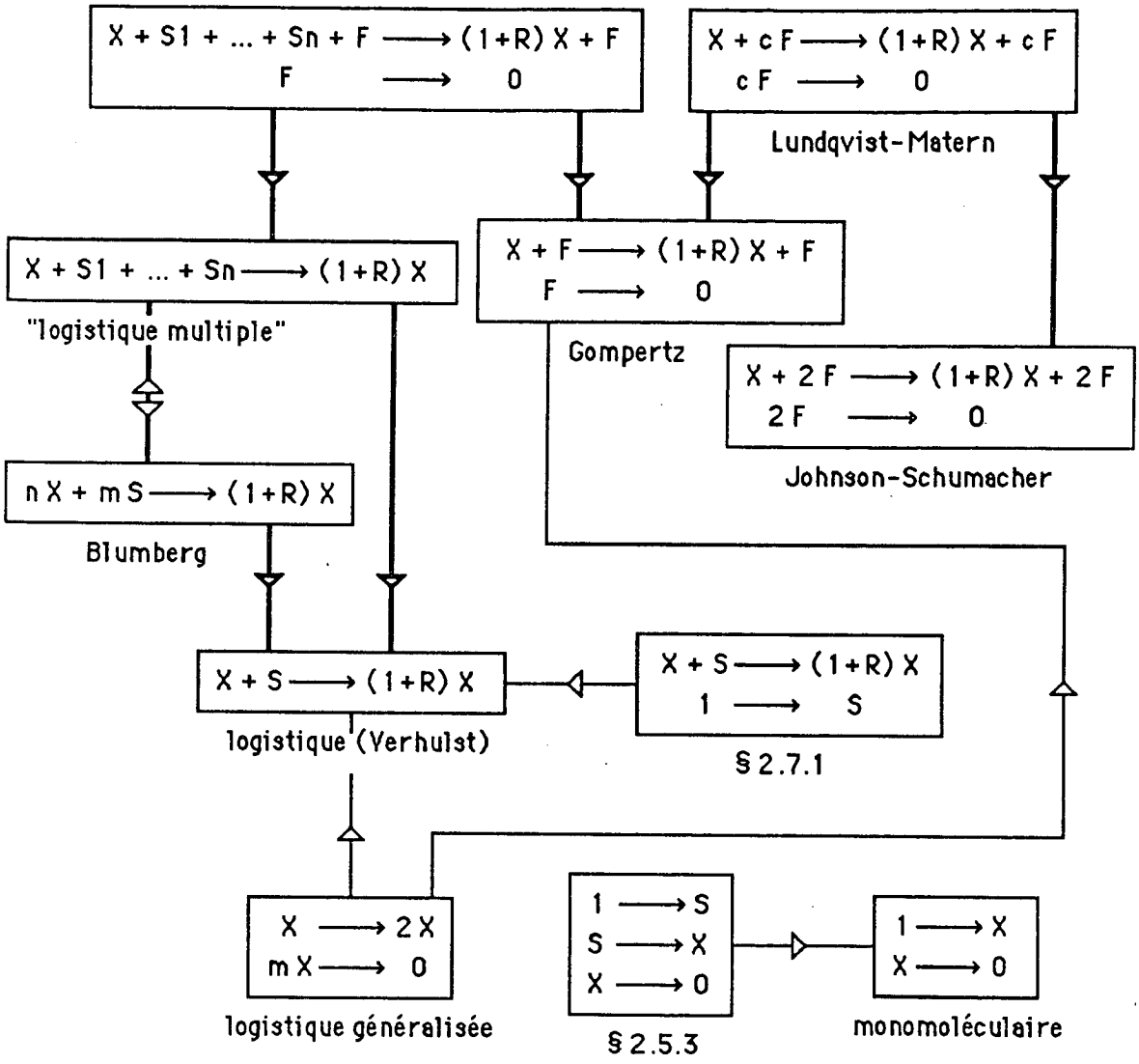
Ce modèle peut par exemple être proposé pour décrire l'action de la lumière.

2.8 - Discussion

Le formalisme de la cinétique chimique permet de proposer des interprétations de nature mécaniste pour des modèles de croissance classiques ou de générer de nouveaux modèles à partir de connaissances biologiques ou de considérations mathématiques. On observe cependant que pour certains types de modèle (cf. § 2.7.2 et 2.4.5), d'ailleurs peu utilisés dans le domaine forestier, ce formalisme doit être complété en introduisant des contraintes.

La contrepartie de la puissance de ce formalisme est qu'il offre souvent "trop" de modèles vraisemblables par rapport à l'état actuel de l'expérimentation ou de l'observation. On peut ainsi générer de multiples courbes de croissance d'allure voisine (asymptote, position du point d'inflexion,...), mais on ne dispose pas de données permettant de choisir entre ces modèles et de valider un mécanisme de croissance (cf. § 1.4). Ces difficultés proviennent en fait de la complexité des phénomènes étudiés et de la nécessaire simplicité des modèles dendrométriques. A ce titre, il est intéressant d'observer que le formalisme chimique permet d'étudier le lien entre des modèles simples et des modèles plus détaillés (Figure 1).

Pour le modélisateur, il s'agit donc d'un outil (parmi d'autres et avec ses limites propres) qui lui permet d'***explorer des hypothèses***, de ***représenter et structurer des connaissances*** souvent qualitatives, et de les ***traduire en langage mathématique***.



relation directe selon l'interprétation pseudo-chimique

relation indirecte (passage à la limite ou forme de l'équation différentielle)

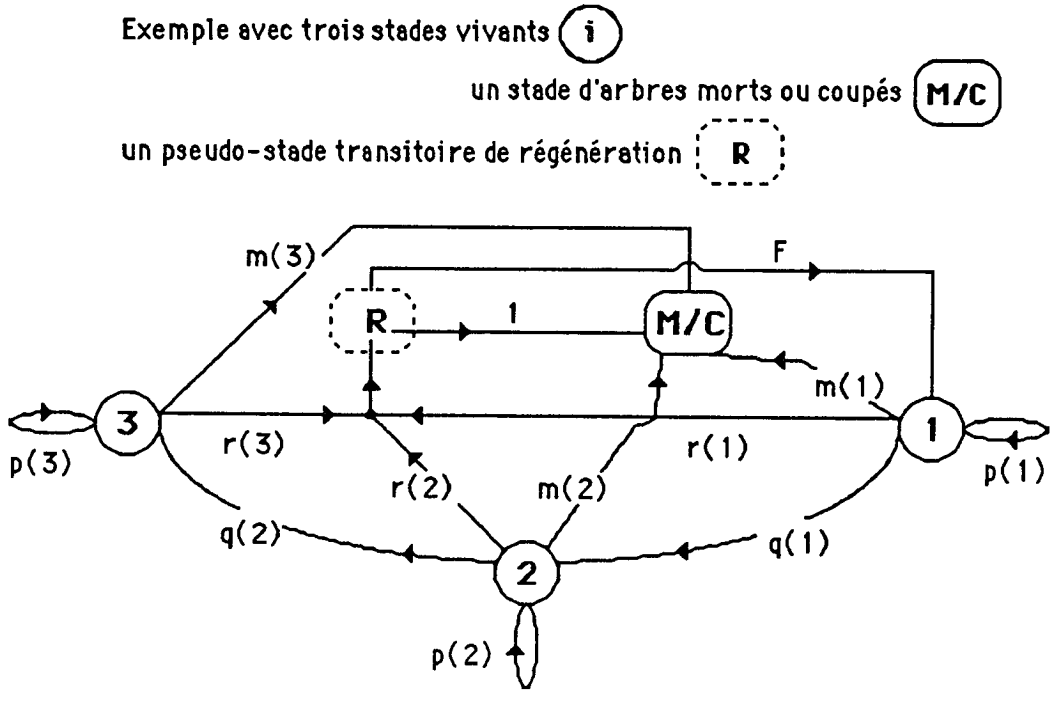
- Figure 1 -

Relations entre différents modèles de courbes de croissance selon leur interprétation de type chimique

3. Construction d'un modèle démographique pour une population structurée en stades

3.1 - Situation du problème et modèle discursif

On considère une population d'individus regroupés en L stades disjoints. On suppose que l'on peut définir sur l'ensemble de ces stades une relation d'ordre total qui correspond à leur succession chronologique du point de vue d'un individu de la population. A un instant donné, chaque individu est ainsi caractérisé par son appartenance à un stade et ***la population est caractérisée par l'effectif des individus dans les différents stades***. On cherche à décrire l'évolution globale de la population et, plus précisément, l'évolution de l'histogramme des stades, noté $(N(t,i))_{i=1,L}$. On suppose de plus que la population est ***isolée***, c'est-à-dire qu'il n'y a pas d'échanges avec l'extérieur.



p(i) survie dans le stade i
q(i) survie avec saut du stade i au stade i+1
m(i) mort ou coupe sans régénération
r(i) mort ou coupe avec régénération
F fécondité d'un arbre qui régénère

- Figure 2 -
Population groupée par stades ou classes
"modèle démographique"

On choisit d'illustrer ici ce type de situations en étudiant une population d'arbres où chaque individu est caractérisé par son appartenance à une classe de diamètre. L'évolution de cette population peut alors être décrite comme suit:

- la dynamique des peuplements forestiers suit des cycles pluriannuels;
- lorsque les arbres grandissent, ils passent progressivement d'une classe de diamètre à la suivante jusqu'à leur mort (naturelle ou par coupe);
- les arbres morts ou coupés peuvent ou non être remplacés par des individus plus jeunes (appartenant à la première classe de diamètre); en toute rigueur, la régénération est fonction des caractéristiques du peuplement auquel appartiennent les arbres morts ou coupés, puisqu'on ne régénère pas des arbres mais des peuplements;
- le domaine spatial étudié est invariant au cours du temps (hypothèse d'isolement de la population).

La dynamique de la population est ainsi caractérisée par l'évolution de l'histogramme des effectifs dans les différentes classes de diamètre et peut être représentée par la figure 2. Cette figure comporte deux types de relations entre classes: des relations concernant la survie ou la mort des individus et des relations décrivant la régénération des individus.

Selon le niveau auquel on se place - arbre, peuplement ou population (ensemble de peuplements disjoints) - on peut construire à partir de ce modèle discursif un modèle stochastique ou un modèle déterministe.

3.2 - Analyse des flux d'arbres et bilan: modèle déterministe

3.2.1 - *Construction du modèle*

On cherche à quantifier la figure 2, en se plaçant au niveau de la population. On analyse ainsi, pour chaque stade, le bilan des flux entrants et sortants. La nature de la croissance forestière conduit à une représentation en temps discret avec un pas de temps Δ, où Δ est un nombre entier d'années, choisi de telle sorte que les arbres ne puissent pas traverser plus d'un stade pendant Δ.

On considère l'ensemble des individus appartenant au stade i à la date t. On note p(i) la proportion de ces individus survivant jusqu'à $t+\Delta$ sans changer de stade et q(i) la proportion passant du stade i au stade i+1. On note enfin m(i) et r(i) les proportions respectives d'arbres qui meurent ou sont coupés entre t et $t+\Delta$ et qui donnent (r(i)) ou non (m(i)) lieu à une régénération; il peut être utile, du point de vue conceptuel, d'introduire un stade supplémentaire qui contient les arbres morts ou coupés. On note enfin f(i) le nombre moyen d'arbres du premier stade générés par des arbres du stade i. On obtient ainsi le modèle:

$$N(t+\Delta) = [U(i,j)]_{i,j} \, . \, N(t) \qquad \text{avec}$$

$$\begin{aligned} U(1,1) &= p(1) + f(1) \, ; \\ U(1,j) &= f(j) \, , \text{ si } j > 1 \, ; \\ U(i,i) &= p(i) \, , \text{ si } i > 1 \, ; \\ U(i,i+1) &= q(i) \, , \text{ si } i < L \, ; \\ U(i,j) &= 0 \, , \text{ sinon} \, . \end{aligned}$$

Le modèle se présente donc comme un modèle ***déterministe, linéaire, récurrent*** (en temps discret) et généralise le modèle de **Leslie** (1945) développé pour les populations groupées par classes d'âge de même amplitude. Ses propriétés, en particulier son comportement asymptotique, ont été étudiées par **Usher** (1969) puis **Houllier** et **Lebreton** (1986).

On peut, par ailleurs, faire l'analogie entre les stades de la population et les compartiments d'un modèle à compartiments (**Pavé**, 1982). On obtient alors le modèle différentiel linéaire: $\frac{dN}{dt} = (U-I).N(t)$. Le choix de l'un ou l'autre des deux modèles comme celui du pas de temps Δ (pour le modèle en temps discret) et de l'amplitude des classes ne sont pas neutres, puisque:
. ils conditionnent la finesse de la description de la dynamique de la population et des résultats, c'est-à-dire le "pouvoir de résolution du modèle";
. ils influent, de façon artificielle, sur les propriétés asymptotiques de ce modèle, en particulier sur le taux de multiplication asymptotique.

A partir du modèle différentiel linéaire, on peut aisément proposer une représentation fonctionnelle de type chimique (cf § 2). Mais dans le cadre particulier d'une population d'arbres, elle ne semble pas porteuse d'une grande signification biologique et elle paraît moins bien adaptée que les représentations graphiques des figures 2 ou 3.

3.2.2 - *Estimation des paramètres du modèle*

L'estimation des paramètres du modèle dépend à la fois de la nature temporelle des données et de l'unité d'échantillonnage (arbre ou placette d'échantillonnage). On distingue ici quelques cas typiques:

(1) On dispose pour ***plusieurs inventaires successifs indépendants***, séparés de Δ, des ***histogrammes globaux*** $N(t_1)$, ..., $N(t_k)$. On peut alors estimer les paramètres de U par régression sur cette série chronologique. Mais cette méthode ne garantit pas que les estimations obtenues vérifient les relations logiques imposées aux paramètres du modèle.
On a par exemple: $0 \leq p(i) \leq 1$ et $0 \leq p(i) + q(i) \leq 1$.

(2) On dispose de ***deux inventaires successifs*** séparés de Δ, avec la ***remesure des mêmes placettes***. Pour chaque placette, on dispose des deux histogrammes successifs. On estime alors U par régression. De la même façon, on n'est pas certain d'obtenir des estimations vérifiant les relations logiques du modèle.

(3) On dispose de ***deux inventaires successifs*** séparés de Δ, avec la ***remesure des mêmes arbres***. On peut alors estimer les paramètres p(i), q(i), r(i) et m(i) en utilisant des distributions multinomiales. On a alors recours, de façon implicite, à un modèle probabiliste qui est explicité au § 2.3.

(4) On dispose d'***un seul inventaire*** pour lequel on a mesuré sur chaque ***arbre*** son diamètre et son ***accroissement*** antérieur. On peut se ramener soit au modèle déterministe en considérant l'accroissement moyen par classe de diamètre, qui détermine un temps de séjour dans chaque stade, soit au modèle probabiliste en reconstituant le diamètre antérieur des arbres.

Dans les cas (1) et (2), on peut choisir de considérer que les coefficients de U sont quelconques (**Lefkovitch**, 1965); on peut alors obtenir des éléments U(i,j) négatifs qui n'ont aucun sens physique ou biologique, mais optimisent le critère utilisé pour l'ajustement. En ce sens, l'approche développée ici pour construire le modèle est du type théorique (elle inclut une certaine connaissance, élémentaire, sur la dynamique de la population) par opposition à l'approche empirique de Lefkovitch. Pour que ces métodes d'estimation soient valides, il faut en fait supposer que:

- pour (1): la dynamique de la population est ***invariante au cours du temps*** (hypothèse d'homogénéité temporelle); cette hypothèse est de plus nécessaire, quelle que soit la méthode d'estimation, si on veut utiliser le modèle à des fins prédictives;

- pour (2), (3) et (4): l'échantillon du premier inventaire est obtenu par ***échantillonnage aléatoire simple*** (des placettes pour (2) et des arbres pour (3) et (4));

- pour (2), (3) et (4), tous les individus échantillonnés (arbres ou placettes) sont ***indépendants*** et ils sont identiques en ce qu'ils suivent le même modèle dynamique (hypothèse d'homogénéité spatiale à rapprocher de son équivalent temporel pour (1)). Le passage à la version stochastique du modèle permet de lever en partie cette troisième hypothèse.

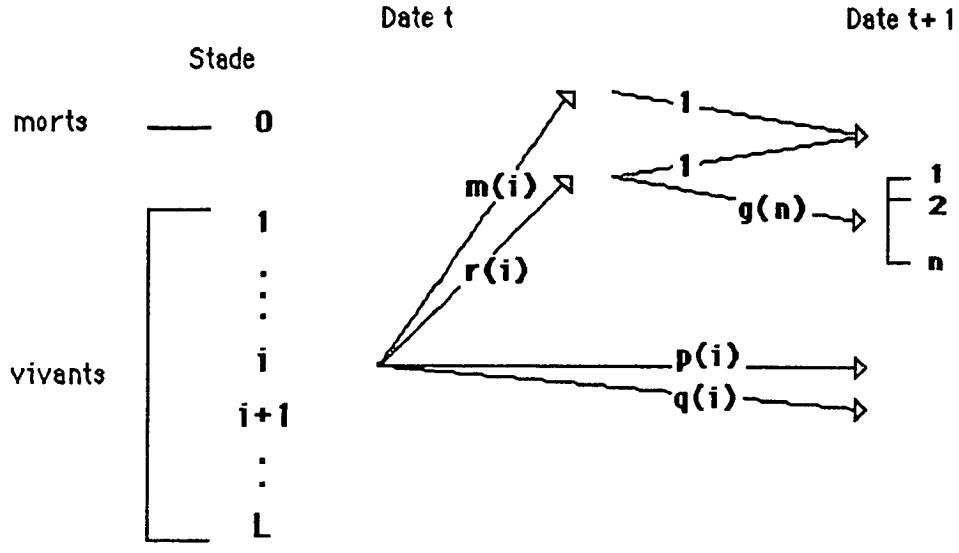
3.3 - Etude de la trajectoire d'un arbre: processus de ramification

Dans la pratique, les arbres ne sont pas tous identiques du fait de leur variabilité génétique propre et de la variabilité (spatiale et temporelle) du milieu. On cherche maintenant à prendre en compte cette variabilité en représentant, par un modèle ***probabiliste***, la trajectoire d'un individu à travers les différents stades successifs possibles (figure 3).

Pour la construction du modèle on se place alors naturellement dans le cadre des ***processus de ramification multitypes*** (**Karlin**, 1966). Comme précédemment, on est amené à décomposer le processus en deux phases:
. une phase de transition pour les individus vivants: passage d'une classe à une autre;
. une phase de remplacement pour les individus morts ou coupés qui donnent ou non lieu à régénération.

La formalisation et le traitement mathématique du modèle font appel aux fonctions génératrices (Figure 3). On montre alors (**Mode**, 1971) que le modèle déterministe proposé au § 2.2.1 peut être obtenu comme l'espérance du processus de ramification conditionnelle à l'histogramme $(N_j)_{j=1,\ldots,L}$.

L'estimation des paramètres du modèle dans le cas de deux inventaires successifs a été envisagée au § 3.2.2 (cas (3)). Dans la mesure où la régénération est essentiellement associée aux coupes pratiquées par les gestionnaires, on est amené à introduire un stade intermédiaire fictif qui regroupe les individus coupés et donnant lieu à régénération. On estime alors, d'une part la proportion r(i) d'individus du stade i, coupés entre t et $t+\Delta$, qui passent par ce stade intermédiaire, et, d'autre part la fécondité apparente de ces individus (de probabilité g(n), cf Figure 3), dont la moyenne est F, c'est-à-dire le rapport du nombre d'individus recrutés sur le nombre d'individus prélevés (Figure 2). On remarque que l'hypothèse d'indépendance des individus, nécessaire aussi bien pour l'estimation des paramètres du modèle que pour son utilisation à des fins prédictives, n'est donc pas strictement respectée.



Fonction génératrice du processus de ramification pour un individu du stade i:

$$\varphi_i(s_1,\ldots,s_L) = p(i).s_i + q(i).s_{i+1} + \sum_{n=0}^{\infty} r(i).g(n).s_1^n$$

Fonction génératrice du processus de ramification pour la population, conditionnelle à l'histogramme $(N_j)_{j=1,..,L}$:

$$\phi(s_1,\ldots,s_L) = \prod_{j=1}^{L} [\varphi_j(s_1,\ldots,s_L)]^{N_j}$$

- Figure 3 -
Processus de ramification multitype pour une population d'arbres groupés par stades

4. Conclusion

Les problèmes rencontrés et les méthodes utilisées ne sont certainement pas spécifiques du domaine forestier; le choix d'exemples appartenant à ce domaine permet de les illustrer et de les situer par rapport à un contexte biologique connu en particulier pour sa complexité. Les trois approches définies dans l'introduction ont ainsi été tour à tour envisagées: la définition d'un schéma fonctionnel pseudo-chimique à partir des mécanismes biologiques ou de la phénoménologie de la croissance relève de l'approche théorique; la recherche a posteriori d'interprétations pour des modèles classiques relève plutôt de l'approche du mathématicien; la construction du modèle démographique est beaucoup plus proche de l'approche empirique, mais on a observé au § 2.2.2 que la méthode de construction du modèle avait imposé certaines contraintes logiques sur les paramètres (méthodes d'estimation (1) et (2)).

Ces limites sont en fait difficiles à tracer et la construction ou le choix de modèles relèvent en général d'allers-et-retours entre ces diverses approches. Ces allers-et-retours ont par exemple lieu, lorsque partant d'une situation et d'un modèle, on se pose le problème de l'estimation des paramètres du modèle: on est alors parfois amené à modifier le modèle et à discuter ses hypothèses (§ 3.2 et 3.3).

La construction ou l'interprétation d'un modèle pour une situation biologique particulière peuvent faire appel à des outils aussi divers que l'analyse, l'algèbre, la théorie des processus stochastiques, ou des représentations fonctionnelles empruntées à la cinétique chimique ou à la théorie des graphes. Ces représentations permettent en particulier de relier efficacement les connaissances biologiques, physiques ou logiques au formalisme mathématique.

Les problèmes d'estimation numérique des paramètres n'ont pas été abordés ici: ils nécessitent, quant à eux, le recours à des méthodes numériques de minimisation, aux théories de l'échantillonnage et de l'inférence statistique.

L'informatique a naturellement sa place au niveau des procédures numériques de simulation ou d'estimation des paramètres. Mais elle peut aussi intervenir pour ce qui relève du calcul formel ou du ***lien entre les représentations fonctionnelles et les modèles mathématiques*** d'une même situation biologique (passage du langage chimique au système différentiel et réciproque, ou passage d'un graphe à un processus de ramification) ou encore des relations de ***"voisinage"*** entre modèles ou entre représentations fonctionnelles.

Ce dernier point mérite une attention particulière puisque ces relations (de voisinage) correspondent aux mécanismes d'inférence par filtrage dans le cadre d'une représentation des connaissances centrée-objet (**Pavé** et **Rechenmann**, 1986) et parce qu'on peut les définir de façon très variée selon en fait l'approche retenue par le modélisateur:

- les modèles de Nelder et de Lundqvist-Matern sont voisins dans la mesure où tous deux ont été utilisés pour décrire la croissance en hauteur des peuplements forestiers;
- les modèles de Gompertz et de Lundqvist-Matern sont par ailleurs voisins en ce sens que tous deux peuvent être interprétés en faisant intervenir un facteur de croissance;
- mais le modèle de Gompertz et le modèle à deux substrats (§ 2.6.1) sont eux aussi voisins puisqu'il est possible d'ajuster le paramètre M2 de façon que le point d'inflexion de la courbe de croissance se situe exactement à $\frac{M}{e}$ (M est l'asymptote de la courbe de croissance) !

Annexe

liste des termes forestiers ou biologiques employés, modèles référencés

peuplement forestier: ensemble d'arbres appartenant à une ou plusieurs espèces, regroupés en un même lieu, soumis à une même gestion, placés dans des conditions écologiques homogènes et interagissant pour leur croissance. Le peuplement est l'unité élémentaire pertinente pour la gestion forestière.

couvert forestier: mesure du degré d'occupation de l'espace aérien par la cime des arbres; un peuplement de couvert fermé intercepte toute la lumière avant qu'elle n'arrive au sol.

substrat: facteur consommé par les arbres pour leur croissance; exemple: éléments minéraux du sol, eau; par extension, la lumière peut être considérée comme un substrat.

facteur de croissance: facteur de type catalytique, nécessaire à la croissance, mais non consommé.

placette d'inventaire: surface unitaire de mesure (de l'ordre de quelques ares) pour les inventaires forestiers.

Modèles:

$$X(t) = p_1\left(1 + p_4 \exp\left(\frac{p_2 - t}{p_3}\right)\right)^{-1/p_4}$$ Nelder § 2.5.4. et 2.4.3.

Monomoléculaires § 2.5.3.
Logistique § 1.3.1. et Pavé même volume

Lundquist-Matern § 2.5.2. $\frac{dX}{dt} = p_2 X\left(\mathrm{Ln}\left(\frac{p_1}{X}\right)\right)^{p_4}$

Johnson-Schumacher § 2.5.2.
Gompertz § 2.5.2.

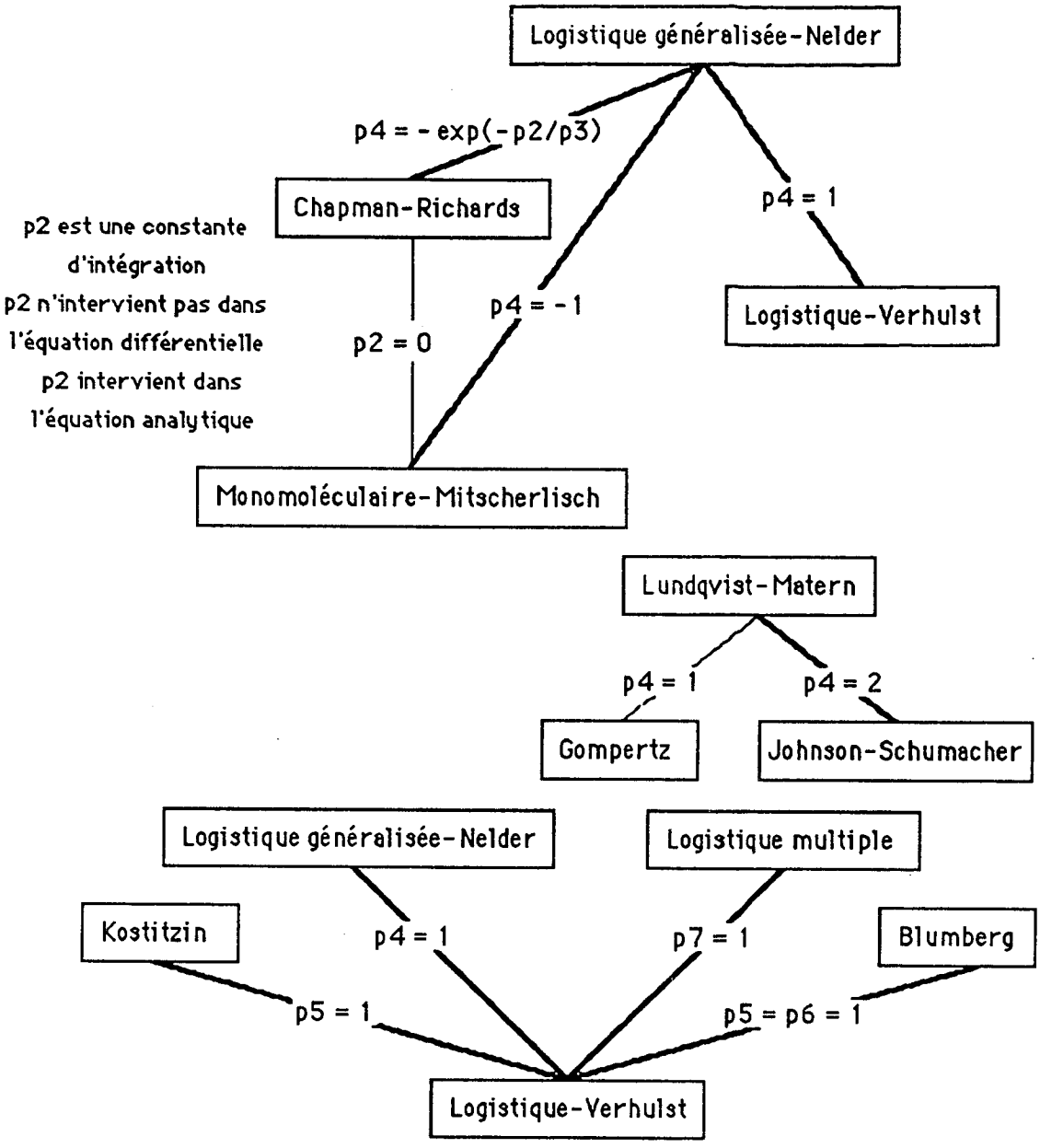
Blumberg § 2.6.1. $\frac{dX}{dt} = p_3 X^{p_5}(M - X)^{p_6}$

Logistique multiple § 2.6.1 $\frac{dX}{dt} = p_3 X \prod_{i=1}^{p_7}(M_i - X)$

Kostitzin Pavé même volume $\frac{dX}{dt} = p_3 X(p_1 - X) - p_5 \int_0^t d\tau$

- Remerciements -

Je remercie Alain BAILLY, Jean BOUCHON, Jean-Luc CHASSE, Jean-Luc GOUZE, Alain PAVE et Dominique PONTIER pour leur collaboration et leurs conseils à un moment ou à un autre de ce travail.



Relations entre modèles par spécialisation-instanciation dans l'équation différentielle

Bibliographie

Bailly A. (1985). Modélisation de la croissance en circonférence du peuplier. Rapport technique de D.E.A., Laboratoire de Biométrie, Université Claude Bernard (Lyon I), 35 pages.

Bertalanffy L. Von (1973). Théorie générale des systèmes. Dunod (Paris), 296 pages.

Chéruy A., C. Gautier et A. Pavé (1982). Analyse des systèmes biologiques: certains aspects méthodologiques liés à la modélisation. In "La notion de système dans les sciences contemporaines", (tome I: méthodologies). Ed. J. Lesourne, Lib. de l'Université, p.75-152.

Debouche C. (1979). Présentation coordonnée de modèles de croissance. Revue de Statistique appliquée, XXVII (4), p.5-22.

Dudek A. et A.R. Ek (1980). A bibliography of worldwide literature on individual tree growth models. College of Forestry, St Paul (Minnesota), Staff papers series 12, 33 pages.

Ek A.R. et A. Dudek (1980). Development of individual tree based stand growth simulator: progress and applications. College of forestry, St Paul (Minnesota), Staff papers series 20, 25 pages.

Emanuel N. et D. Knorre (1975). Cinétique chimique. Editions MIR, 448 pages.

Garfinkel D., J.D. Rutledge et J.J. Higgins (1962). Simulation and analysis of biochemical systems: I representation of chemical kinetics. Commun. Assoc. Computing Machinery, 4, p.559-562.

Garfinkel D. (1968). A machine independent language for the simulation of complex chemical and biochemical systems. Computers Biomed. Res., 2, p.31-44.

Garcia O. (1983). A stochastic differential model for the height growth of forets stands. Biometrics, 39, p.1059-1072.

Harper J.L (1977). Population biology of plants. Academic Press (London), 2nd ed., 892 pages.

Houllier F. et J.D. Lebreton (1986). A renewal equation approach to the dynamics of stage-grouped populations. Mathematical Biosciences (à paraître).

Karlin S. (1966). A first course in stochastic processes. Academic Press (New York), 502 pages.

Lebreton J.D. et C. Millier (1982). Modèles dynamiques déterministes définis par des équations différentielles. In Modèles dynamiques déterministes en biologie, Lebreton et Millier, Masson (Paris), p.13-58.

Lefkovitch L.P. (1965). The study of population growth in organisms grouped by stages. Biometrics, 21, p.1-18.

Legay J.M. (1973). La méthode des modèles, état actuel de la méthode expérimentale. Informatique et Biosphère (Paris), p.7-73.

Leslie P.H. (1945). On the use of matrices in population mathematics. Biometrika, 33, p.213-245.

Maugé J.P. (1975). Modèles de croissance et de production des peuplements de Pin maritime. Publication A.FO.CEL., p.227-249.

Millier C. (1982). Courbes de réponse. In Modèles dynamiques déterministes en biologie, Lebreton et Millier, Masson (Paris), p.151-170.

Mode C.J. (1971). Multitype branching processes, theory and applications. Elsevier (Londres), 330 pages.

Munro D. (1974). Forest growth models: a prognosis. In Growth models for tree and stand simulation, J. Fries (ed.), Royal College of Forestry (Stockholm), Res. Notes 70, p. 7-21.

Pavé A. et Pagnotte Y. (1977). An approach to computer-aided design: a tool for mathematical modelling in biology and ecology. Comput. Biol. Med., 7, p.301-310.

Pavé A. (1980). Contribution à la théorie et à la pratique des modèles mathématiques pour l'analyse dynamique des systèmes biologiques. Thèse de Doctorat ès Sciences, Université Claude Bernard (Lyon I), 147 pages.

Pavé A. (1982). Modèles à compartiments linéaires. In Modèles dynamiques déterministes en biologie, Lebreton et Millier, Masson (Paris), p.99-134.

Pavé A. et F. Rechenmann (1986). Computer aided modelling in biology: an artificial intelligence approach. In Artificial Intelligence and Simulation, SCS Rev.

Pavé A. Interprétation et construction de modèles de la dynamique des populations à l'aide de schémas fonctionnels (même volume p. 49-82).

Pienaar L.V. et K.J. Turnbull (1973). The Chapman-Richards generalization of Von Bertalanffy's growth model for basal area growth and yield in even-aged stands. Forest Science, 19 (1), p.2-

Usher M.B. (1969). A matrix model for forest management. Biometrics, 25, p.309-315.

Vila J.P. (1982). Méthodes d'identification des modèles dynamiques. In Modèles dynamiques déterministes en biologie, Lebreton et Millier, Masson (Paris), p.171-195.

LA LOI EXPONENTIELLE ET SES VERIFICATIONS EXPERIMENTALES EN BIOLOGIE

Jean-Luc GOUZÉ
INRIA Sophia-Antipolis
06560 Valbonne

Antoine SCIANDRA
Station d'écologie marine
06230 Villefranche-sur-mer

Nous examinons ici quelques problèmes liés à l'utilisation de la loi exponentielle en biologie.

Nous nous plaçons dans le cadre du formalisme continu des équations différentielles. La loi exponentielle est alors solution de l'équation:

$$x'(t) = k\, x(t)$$

x(.) est une fonction réelle du temps t, avec $x(0) = x_0$.
x' désigne la dérivée par rapport au temps.
k est une constante réelle.

1. Conditions de validité

Elles peuvent se formuler ainsi: la quantité physique dont on observe l'évolution est bien représentée par une fonction réelle continue x(t) du temps; son taux d'accroissement (rapport de la dérivée x'(t) à la variable x(t)) est constant (indépendant du temps et de la valeur de x).

Prenons un exemple, emprunté à la physique, mais qui pourra être aussi interprété comme un processus de "naissance et de décès" biologique,et qui permet de cerner les hypothèses (Pielou, 1977, Feller 1940a). Considérons, à l'instant t=0, N_0 atomes (ou individus) pouvant se désintégrer (ou mourir) à chaque instant t.

Nous souhaitons prédire l'évolution N(t) de la population. Les hypothèses classiques sont:

(i) on suppose que les effectifs N(t), qui sont des entiers, sont très grands, et sont "bien représentés" par une fonction continue réelle x(t),

(ii) on suppose que $k = -\frac{x'(t)}{x(t)}$, k taux d'accroissement de x, est une constante réelle (positive dans notre cas), indépendante de x.

On obtient alors le résultat bien connu:

$$x(t) = x_0\, e^{-k\,t} = N_0\, e^{-k\,t}$$

On peut faire quelques remarques sur l'hypothèse (i):
- elle est en général invérifiable: on ne connait pas l'erreur faite en confondant x(t) et N(t). Le terme "bien représenté " est sujet à interprétation (voir, pour une discussion sur la pertinence du formalisme continu et différentiel, Lobry (1985)),
- la notion d'ordre de grandeur demanderait à être précisée.

Cependant, dans ce cas simple, on peut se livrer à une autre modélisation du phénomène qui permet des comparaisons intéressantes: on modélise la désintégration d'un atome par un processus stochastique, et on suppose que la probabilité de désintégration pendant un petit intervalle Δt est constante et égale à $(k.\Delta t)$ plus des termes d'ordre supérieur en Δt.

Un calcul simple donne alors la probabilité d'avoir n individus à l'instant t (Feller, 1940a):

$$p_n(t) = \binom{N_0}{n} e^{-\lambda n t}(1 - e^{-\lambda t})^{N_0 - n}$$

Si on calcule l'espérance M(t) de ce processus, on trouve que M(t) vérifie l'équation:

$$M'(t) = -k\,M(t) \quad \text{avec} \quad M(0) = N_0$$

et donc que la fonction M est égale à la fonction x de la modélisation précédente.

On peut avoir une bonne estimation de l'erreur relative faite en confondant le processus et son espérance, par le rapport de l'écart-type et de l'espérance:

$$\frac{\sigma(t)}{E(t)} = \frac{(N_0\, e^{-\lambda t}(1 - e^{-\lambda t}))^{1/2}}{N_0\, e^{-\lambda t}} = \left(\frac{e^{\lambda t} - 1}{N_0}\right)^{1/2}$$

Ce rapport:

- devient grand de manière exponentielle quand le temps augmente (et que le nombre d'atomes restants diminue).
- à t fixé, varie comme $(N_0)^{-1/2}$.

2. Emploi en biologie

La loi exponentielle est très employée, et souvent sans justifications *a priori* d'ordre biologique.

L'hypothèse (i) de bonne représentation est en général invérifiable. Le biologiste utilise souvent des concentrations comme fonction x(t), ce qui complique encore les choses, puisque le volume considéré entre en compte.

L'hypothèse (ii) de constance du taux d'accroissement peut être valide dans un certain domaine; elle devient en général fausse à partir du moment où il y a interférence entre les individus, par exemple compétion pour un substrat. Remarquons que la loi peut donc devenir fausse quand il y a beaucoup d 'individus, et que cette tendance s'oppose à la validité de la représentation par un processus stochastique du paragraphe précédent (la représentation est d'autant meilleure qu'il y a plus d'individus).

L'emploi de la loi exponentielle est aussi renforcé par d'autres raisons:

- la simplicité de cette loi, qui permet de résumer une courbe par un seul coefficient,
- son utilisation fréquente et fructueuse (mais plus argumentée) en physique,
- sa formulation explicite (pour d'autres lois on ne peut raisonner que sur l'équation différentielle),
- sa facilité d'ajustement, point sur lequel nous allons insister.

En effet la loi exponentielle dépend d'un seul paramètre, et peut être linéarisée par la transformation:

$$y(t) = \ln x(t)$$

qui donne:

$$y(t) = y_0 - k\,t$$

On peut alors estimer k (et éventuellement y_0) par régression linéaire.

Les deux raisons (formulation explicite et linéarisation possible) ont puissamment contribué au succès de cette loi, surtout à l'époque où les moyens de calcul informatiques n'étaient pas développés.

On sait maintenant que l'ajustement sur la transformation logarithmique (dite "par les log") introduit un biais qui peut être important (voir ci-après).

3. Essais d'ajustement

Nous avons extrait un jeu de données expérimentales de la littérature biologique (Tande and Bämsted, 1985). Ces données mesurent la décroissance en fonction du temps du contenu de l'estomac d'un copépode (mesure par fluorescence de la quantité de pigments chlorophylliens). Elles sont assez représentatives des données biologiques usuelles, tant par leur nombre que par la dispersion des mesures.

Nous avons effectué un ajustement pour plusieurs lois, dont la loi exponentielle avec ou sans transformation logarithmique. Le critère utilisé est celui des moindres carrés ordinaires, le plus fréquemment employé en biologie.

La méthode de minimisation utilisée est une méthode de quasi-Newton, basée sur un programme NAG E04JBF, et couplée à une méthode robuste d'intégration numérique. Pour chaque valeur des paramètres on calcule le critère en intégrant l'équation depuis la condition initiale, qui peut aussi être un paramètre.

Les lois choisies pour illustration peuvent toutes avoir une justification biologique (Lebreton et Millier, 1982). On utilise:

- la loi de Hill, applicable par exemple dans le cas d'enzymes allostériques, qui s'écrit:

$$x' = -\,a\,\frac{x^n}{b^n + x^n}\,x$$

avec n fixé ici à la valeur raisonnable de 2 (voir Monod et al., 1965, pour des exemples).

- une loi quadratique:

$$x' = -\,a\,x^2$$

- une loi exponentielle avec seuil:

$$x' = -\,a\,(x - b)$$

dont la solution admet la droite x=b comme asymptote, au lieu de la solution nulle comme pour les lois précédentes.

- enfin, à titre d'amusement, une loi de type proie-prédateur dont on n'observerait que l'une des variables d'état, par exemple la proie,

$$x' = a\,x - b\,x\,y$$
$$y' = b\,x\,y - c\,y$$

On observe x(t); on sait que les solutions x(t) et y(t), pour toute valeur de la condition initiale, oscillent périodiquement (Pielou, 1977).

Le tableau ci-dessous compare le critère (somme des carrés des écarts) pour les différentes lois; une comparaison plus visuelle est donnée par des courbes représentant les points expérimentaux et les résultats donnés par le modèle:

Loi	Critère	Paramètres identifiés	Figure
exponentielle	7.04	$k = 0.056$ $x_0 = 12.0$	1
exponentielle (reg.log)	22.2	$k = 0.0014$ $x_0 = e^{2.5}$	
Hill	4.55	$a = 0.097$ $b = 1.75$ $x_0 = 23.5$	2
quadratique	7.75	$a = 0.0137$ $x_0 = 50.$	3
exp. (seuil)	6.31	$a = 0.061$ $b = 0.1$ $x_0 = 13.01$	4
prédateur proie	5.12	$a = 0.005$ $b = 0.00277$ $c = 0.0294$ $x_0 = 20.,\ y_0 = 30.$	5

Dans tous les cas, on obtient une valeur du critère inférieure (ou comparable pour la loi quadratique) à celle du critère de la loi exponentielle. De plus, l'ajustement *de visu* est comparable, sinon meilleur. On peut remarquer le biais important introduit par l'ajustement sur la transformation logarithmique.

4. Conclusion

On peut conclure sur deux points pratiques:

- il est important d'avoir des données sur un intervalle de temps suffisamment long pour bien discriminer les différentes lois. Beaucoup d'expériences biologiques s'arrêtent très tôt, dès que les mesures tendent grossièrement vers zéro, avec l'hypothèse implicite que les mesures suivantes tendront aussi régulièrement vers zéro,

- l'ajustement par courbes semi-log devrait maintenant être évité.

Et un point méthodologique:

- la qualité de l'ajustement est difficile à apprécier numériquement par un critère, et ne suffit pas à valider un modèle. Cette remarque avait déjà été utilisée par Feller (1940b)

pour critiquer l'ajustement de la loi logistique à des données expérimentales. Il paraît donc important de choisir un modèle correspondant à des considérations biologiques a-priori précises. Sinon, la loi exponentielle n'a pour seul argument en sa faveur que sa simplicité d'écriture et de représentation, qui peut masquer des concepts importants comme celui de seuil (loi de Hill) ou d'autolimitation (loi en x^2). De plus, il y a d'autres lois (celle en x^2) qui ne dépendent que d'un paramètre tout aussi significatif.

Certes, il est vrai (et évident mathématiquement) que cette loi approxime correctement **toutes** les lois sur un certain intervalle (c'est l'approximation au premier ordre par linéarisation). Cette propriété cependant est locale, et donne donc en général très peu de renseignements sur le modéle global.

Enfin, une remarque d'ordre épistémologique:

- nous avons examiné sept articles plus ou moins récents relatant la même expérience. Quatre d'entre eux, pour justifier l'emploi de la loi exponentielle, citent les deux articles les plus anciens. L'un de ceux-ci (Mackas and Bohrer, 1976) contient, en légende d'une figure où sont tracées les mesures pour une expérience de remplissage puis de vidage de l'estomac: " the lines shown are statistical best fits for increasing (linear) and decreasing (exponential) values of G" (G est le contenu stomacal). C'est la seule référence à la loi exponentielle dans tout l'article. L'autre (**Gauld**, 1957) contient de belles photos de copépodes, mais pas de mesures quantitatives ni d'allusion à une loi de vidage de l'estomac. La loi exponentielle, dans ce contexte, est donc employée sans justification *a priori*, plutôt pour rendre compte du fait que la décroissance n'est pas linéaire.

Bibliographie

Feller W. (1940a) Volterra's theory of struggle for existence treated probalistically, in "Applicable mathematics of non-physical phenomena "Oliveira-Pinto and Conolly ed (1982), Ellis Horwood limited.

Feller W. (1940b) On the logistic law of growth and its empirical verifications in biology, in "Applicable mathematics of non-physical phenomena " Oliveira-Pinto and Conolly ed. (1982), Ellis Horwood limited.

Gauld D.T. (1957) A peritrophic membrane in calanoïd copepods. Nature 4554, vol 179,325-326.

Lebreton J.D, Millier C., ed (1982) Modèles dynamiques déterministes en biologie. Masson.

Lobry C. (1985) Remarques sur l'utilisation des systèmes différentiels ailleurs qu'en physique (preprint)

Mackas D.,Bohrer R. (1976) Journal of experimental marine biology and ecology. 25,77-85.

Monod J., Wyman J., Changeux J.P. (1965) On the nature of allosteric transitions: a plausible model. J. Mol. Biol. 12, 88-11.

Pielou E.C.(1977) Mathematical Ecology. Wiley.

Tande K.S. and Bämstedt U.(1985) Grazing rates of the copepods Calanus glacialis and C. finmarchicus in arctic waters of the Barents Sea. Marine Biology 87, 251-258.

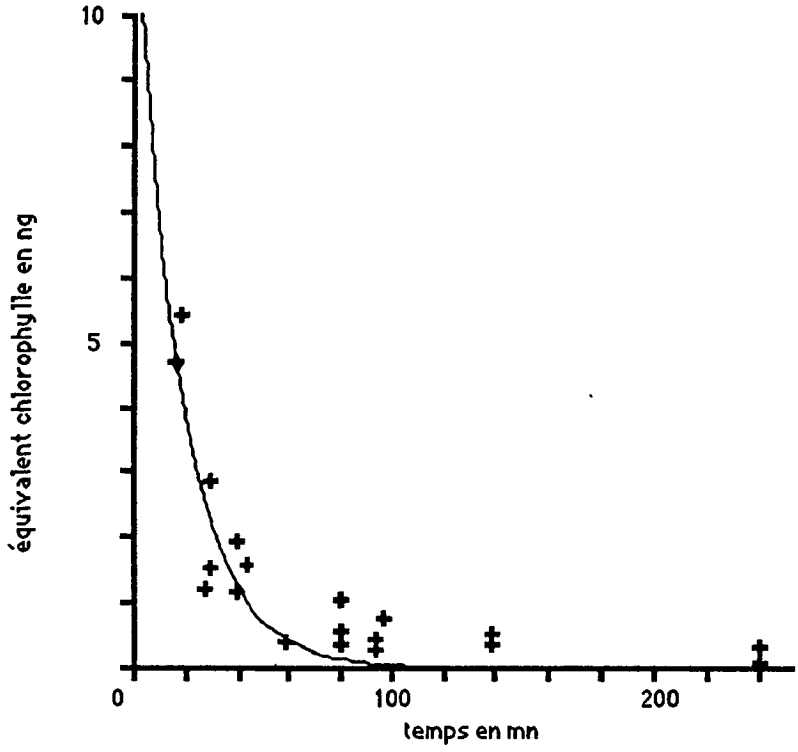


Figure 1 - Contenu de l'estomac en fonction du temps, pour une loi exponentielle.

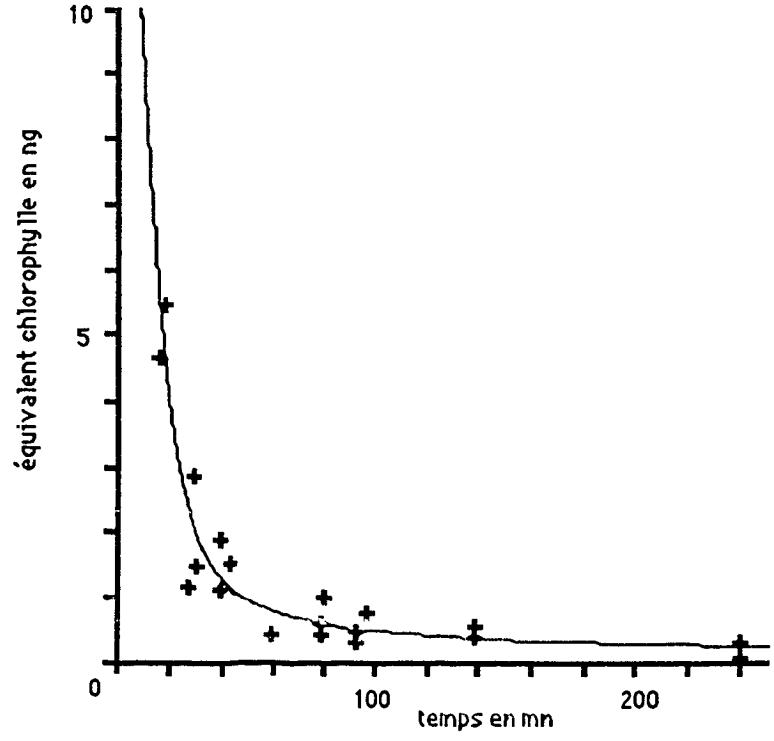


Figure 2 - Contenu de l'estomac en fonction du temps, pour une loi de Hill.

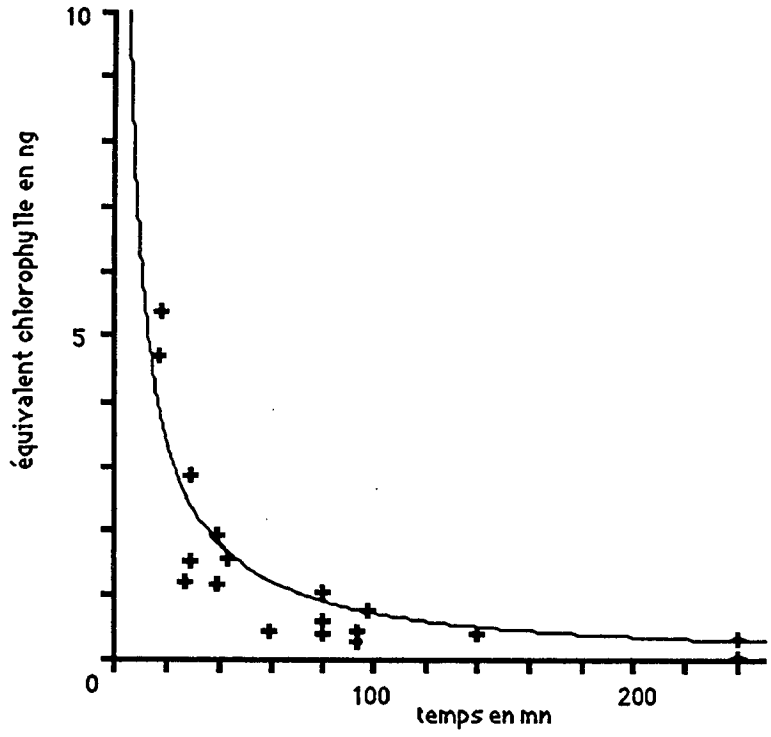


Figure 3 - Contenu de l'estomac en fonction du temps, pour une loi quadratique.

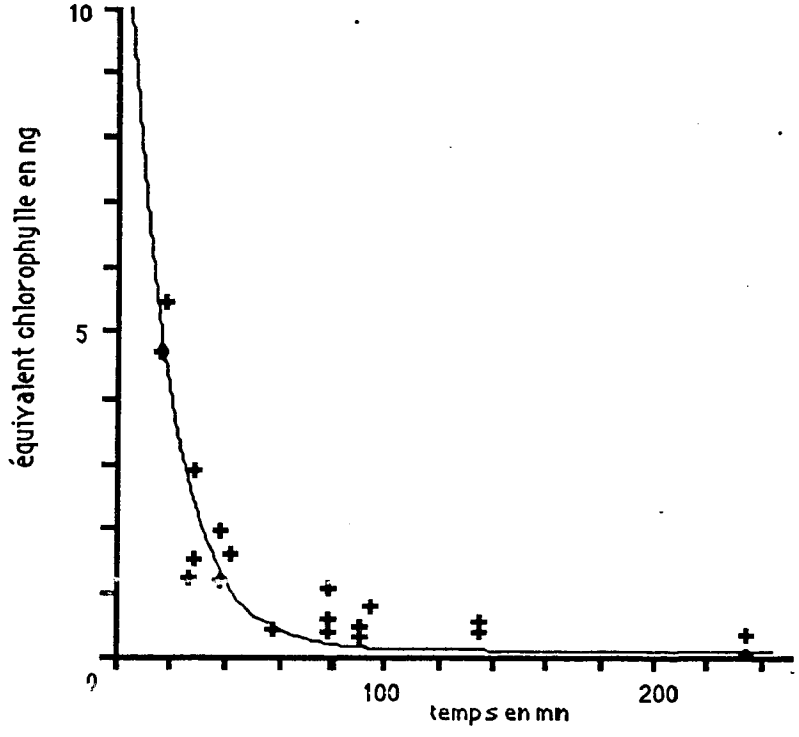


Figure 4 - Contenu de l'estomac en fonction du temps, pour une loi exponentielle avec seuil.

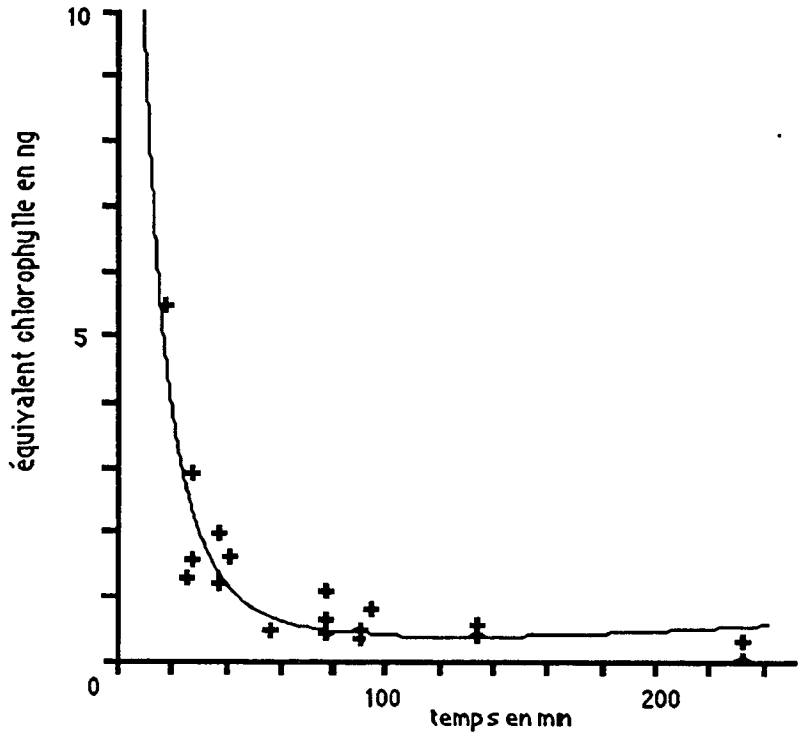


Figure 5 - Contenu de l'estomac en fonction du temps, pour une loi prédateur-proie, on observe x.

MODÉLISATION ET IDENTIFICATION EN AUTOMATIQUE
Transferts possibles vers les bio-systèmes

Sylviane GENTIL
Laboratoire d'Automatique de Grenoble UA 228
ENSIEG-INPG - B.P. 46
38402 Saint Martin d'Hères

I - Introduction

Nous restreindrons dans ce qui suit la notion d'automatique à celle d'analyse et de conception de commande pour des *systèmes continus*, c'est-à-dire des systèmes pouvant être décrits par des équations différentielles, ordinaires ou aux dérivées partielles. L'automaticien est confronté depuis toujours à la notion de commande de tels systèmes: faire en sorte que le système se comporte comme il le désire, que les grandeurs qui le caractérisent restent dans des limites préfixées ou éventuellement conservent une valeur très particulière, malgré d'éventuelles perturbations prévisibles ou non, mesurables ou non. Pour atteindre cet objectif, il a besoin d'une représentation du procédé à commander: si dans l'approche dite "classique" cette représentation peut se contenter d'être graphique (réponses types, d'où les différents diagrammes et abaques de synthèse indicielle ou fréquentielle), il n'en est plus question lorsque le procédé devient un tant soit peu complexe. Dans l'approche dite "moderne", le procédé est toujours représenté par un *modèle mathématique*, sur lequel on effectue les calculs d'analyse, de stabilisation ou d'optimisation.

Le modèle mathématique représente donc un système, défini par rapport à un environnement; cet environnement agit sur le système (entrées de commande si l'on peut les manipuler, de perturbations dans le cas contraire) et reçoit son influence (sorties, qui représentent généralement les grandeurs à réguler).

Deux voies ont été explorées pour construire le modèle mathématique d'un système. L'une, qualifiée de *modélisation*, est en fait la démarche classique des sciences exactes: on intègre toute la connaissance que l'on a des divers processus élémentaires, on les relie entre eux, on fait clairement ressortir le corps des hypothèses nécessaires à la représentation que l'on élabore.

On parle alors de *modèle de connaissance*, et l'on attend du modèle qu'il se comporte comme le procédé dans les situations les plus variées. Cette approche est parfois qualifiée de mécaniste dans des disciplines comme la biologie ou l'écologie. Le modèle ainsi obtenu est souvent complexe (comportant beaucoup de variables et de paramètres, non linéaire...). Il peut être utilisé surtout pour la *simulation* du système.

L'autre approche est une approche expérimentale: ayant observé le comportement du système dans une ou plusieurs occasions, autrement dit ayant enregistré les entrées u(t) et les sorties y(t), on désire en tirer *directement* la relation $y(t) = f(u(t), t)$. On obtient ainsi un *modèle de comportement.*

Cette approche serait utopique si on ne lui adjoignait pas des hypothèses très restrictives sur la fonction f: par exemple, si le modèle ne doit servir qu'au calcul de la commande du système, il suffit qu'il représente correctement son fonctionnement normal (gamme de variations d'entrées faible). *L'identification* des systèmes consiste en ce

processus d'expérimentation, de restriction du modèle à une classe très particulière (très souvent, pour la commande du procédé, on cherche une représentation linéaire par équation différentielle ou aux différences), suivie d'une estimation des paramètres entrant dans la relation cherchée.

Bien sûr les deux approches mentionnées ci-dessus sont plus complémentaires que concurrentes. Bien souvent on identifie un système, on calcule la commande sur ce modèle simplifié et on l'essaye en simulation sur un modèle plus complexe avant de l'appliquer sur le système réel.

Dans le cadre de l'identification, on essaye de maximiser l'information que l'on peut obtenir d'une seule expérimentation: cela conduit à des problèmes non triviaux qui sortent du cadre de cet exposé; pour rester simple, soulignons que si l'entrée d'un système n'évolue pas, sa sortie n'évolue pas non plus et la seule relation que l'on peut obtenir entre les deux est une constante. Donc l'entrée doit "bouger beaucoup" pour qu'une bonne relation entrée-sortie soit identifiée.

Dans le cadre de la modélisation des procédés mal connus (ceci s'applique donc tout particulièrement aux systèmes biologiques ou écologiques), le nombre d'hypothèses hasardeuses sur les relations mathématiques entre variables est grand et les paramètres quantifiant ces relations impossibles à mesurer directement. On résout dans ce cas, comme en identification, un problème d'estimation de paramètres à partir de données expérimentales. En principe, ce problème peut trouver une solution dans le cadre de l'optimisation non linéaire qui permet de minimiser toute fonctionnelle de l'écart entre le comportement du modèle et l'expérience. Toutefois, la structure des équations est imposée par des considérations autres que mathématiques, et peut être relativement complexe: dans ce cas, le problème posé peut très bien avoir une infinité de solutions, ce qui est bien plus dangereux que s'il n'en avait aucune. Ceci conduit au problème très intéressant de *l'identifiabilité* des modèles, qui sort lui aussi du cadre de cet exposé, mais qui commence à être bien défini, sinon bien résolu (Walter, 1984).

Dans le cadre du Laboratoire d'Automatique de Grenoble, de nombreux problèmes de modélisation ont été abordés: dans les industries pétrochimiques, papetières ou pharmaceutiques, dans le cadre de systèmes biomédicaux ou environnementaux. D'autres études ont porté sur l'identification des systèmes linéaires. De cette dernière activité est né un progiciel, depuis peu commercialisé, qui va être décrit dans ce qui suit.

L'identification des systèmes linéaires pouvant paraître très éloignée des préoccupations des biologistes ou des écologistes, nous nous efforcerons de montrer ensuite les liens que l'on peut établir entre les deux disciplines.

II - Environnement général

Le logiciel d'identification que l'on présente ici peut être considéré comme un tout ou bien s'insérer lui-même dans un environnement plus général constituant un système complet de CAO pour l'automatique. Il s'agit du progiciel SIRENA [Yem et al., 1984] qui constitue une nouvelle génération d'outils de simulation, mettant à la disposition de l'utilisateur les moyens classiques de simulation utilisés en automatique (systèmes continus, discrets et échantillonnés) mais aussi un ensemble, évolutif, de techniques modernes, comme l'estimation de paramètres.

La communication homme-machine est réalisée ici sous forme d'un dialogue auto-guidé permettant pour l'expert des enchaînements abrégés de directives à concurrence d'un contexte détecté comme ambigu, et offrant au débutant un dialogue question-réponse lui facilitant un apprentissage rapide.

SIRENA, c'est à la fois une base de données et un ensemble de logiciels de calculs gérés d'une façon transparente, sur le plan informatique, par l'utilisateur. Voici par exemple une configuration simple de commande d'un moteur déclarée sous SIRENA (figure 1).

```
⇒ DECL-SYSTEME CONT SI E POS

  TYPE          : CONTINU
  NOM           : SI
  ENTREE (S)          : E
  SORTIE (S)    : POS

  POS           = INT * VIT
  VIT           = MOT * X
  X             = REL (U)
  U             = E – POS – TAC * VIT
```

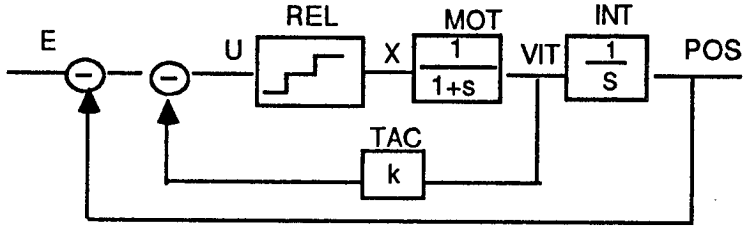


Figure 1 - Commande de moteur.

Cette description étant faite, l'utilisateur décrit et paramétrise (numériquement ou formellement) les différents blocs de son schéma. La simulation proprement dite nécessite ensuite la connaissance d'une entrée. SIRENA reconnait dans ce cadre un certain nombre d'entrées-type (échelon, séquence binaire, rampe, etc.) mais donne également accès à des entrées de type fichier utilisateur. A l'issue de la simulation, une visualisation des signaux mis en jeu est activée avec une présentation standard. La présence d'un éditeur graphique intégré au système, offre alors la possibilité de gérer à sa guise le ou les signaux visualisés (zoom, effacement, superposition, déplacements de courbes, légendes, etc.). Dans un contexte d'identification, l'utilisateur suit une démarche déclarative analogue à la précédente (nom du modèle, noms des entrées et sorties), mais cette fois ce sont les paramètres d'un "bloc modèle" que l'on cherche à déterminer, les signaux entrées-sorties étant déjà connus. Ces derniers peuvent soit découler d'une phase précédente de simulation (cas où l'on voudrait simplifier un modèle complexe par exemple), soit être constitués par des fichiers utilisateurs (mesures effectuées sur un système réel).

III - Les méthodes d'identification mises en œuvre [Foulard et al., 1986]

Les méthodes d'identification diffèrent entre elles essentiellement par le modèle sur lequel on peut les appliquer, et en conséquence le type de mesures effectuées sur le procédé, et le critère d'évaluation de la ressemblance entre le comportement du système et celui du modèle.

Actuellement le système d'identification s'adresse à des modèles dynamiques linéaires mono ou multivariables, soit de type continu, dans le domaine fréquentiel, soit de

type discret dans le domaine temporel, (la levée de cette restriction type-domaine fait partie des développements en cours d'intégration). On supposera par la suite que ces données ont déjà été "préparées" en vue de l'identification, on entend par là, élimination de dérive, centrage, filtrage, élimination des points aberrants, interpolation de données manquantes, etc. Pour cela l'utilisateur dispose dans SIRENA de directives ad-hoc sous la rubrique "Traitement-Signaux".

Deux types de critères ont été retenus: la minimisation d'une erreur de prédiction et la minimisation d'une erreur de sortie. Dans la première catégorie on trouve la méthode des Moindres Carrés Simples (MCS), la méthode des Moindres Carrés Etendus (MCE) et la méthode du Maximum de Vraisemblance (MV); toutes trois s'appliquent à la détermination d'un modèle discret dans la domaine temporel. Dans la deuxième catégorie, est implémentée la Méthode du Modèle (MM) à la fois dans le domaine temporel et dans le domaine fréquentiel. Nous allons maintenant préciser dans ce qui suit les principes théoriques sur lesquels reposent les méthodes employées.

Dans le but de simplifier les notations, on formalisera le problème dans un cadre monovariable. L'extension multi-entrées est triviale; quant aux systèmes également multi-sorties, ils sont considérés comme un ensemble de sous-systèmes mono-sortie.

1) Minimisation d'une erreur de prédiction

La représentation du système est donnée figure 2, la suite (u_t, y_t) représente l'ensemble des couples entrées-sorties mesurés à des instants d'échantillonnage réguliers. Bruits de mesures, écarts modèle-procédé sont globalement représentés par la séquence (b_t) modélisée par un bruit blanc (ε_t) filtré.

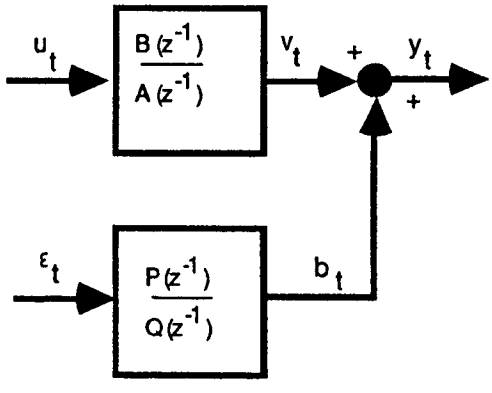


Figure 2 - Représentation du système.

On a noté A, B, P et Q des polynômes en z^{-1} de degrés respectifs n, m, p et q. Il vient alors:

$$A(z^{-1})\, v_t = B(z^{-1})\, u_t$$

$$v_t + a_1\, v_{t-1} + \ldots + a_n\, v_{t-n} = b_0\, u_t \ldots + \ldots + b_m\, u_{t-m} \qquad (1)$$

$$Q(z^{-1})\, b_t = P(z^{-1})\, \varepsilon_t$$

$$b_t + \alpha_1\, b_{t-1} + \ldots + \alpha_q\, b_{t-q} = \varepsilon_t + \ldots + \beta_p\, \varepsilon_{t-p} \qquad (2)$$

D'où la relation générale du type prédiction:

$$A(z^{-1})\, y_t = B(z^{-1})\, u_t + \frac{A(z^{-1})\, P(z^{-1})}{Q(z^{-1})}\, \varepsilon_t \qquad (3)$$

1.1) La méthode des Moindres Carrés Simples (MCS)

Cette méthode consiste à réduire la forme générale (3) à:

$$A(z^{-1})\, y_t = B(z^{-1})\, u_t + \varepsilon_t \qquad (4)$$

C'est-à-dire à poser $A(z^{-1})\,.\,P(z^{-1}) / Q(z^{-1}) \equiv 1$ La quantité ε_t devient alors l'erreur d'équation dont on aura à minimiser la norme euclidienne sur l'horizon de mesure. C'est un problème quadratique par rapport aux inconnues a_i, b_j dont la solution explicite est donnée par :

$$\underline{\theta}_{MCS} = (X^T X)^{-1}\, X^T \underline{y} \qquad (5)$$

avec les notations:

$$\underline{\theta}^T = [b_o, ..., b_m, a_1, ..., a_n]$$

$$\underline{z_t}^T = [u_t, ..., u_{t-m}, -y_{t-1}, ..., -y_{t-n}] \qquad (6)$$

$$\varepsilon_t = y_t - \underline{z_t}^T.\underline{\theta}$$

$$t = n+1, ..., N$$

et $$X^T = [\underline{z}_{n+1}, ..., \underline{z}_N] \qquad (7)$$

L'avantage de cet estimateur est la simplicité. Cependant l'hypothèse ε_t, blanc, est peu réaliste en raison de la structure du modèle de bruit implicitement contenu dans (4) qui implique un bruit de sortie b_t filtré par les pôles du système. On sait que dans ces conditions $\underline{\theta}$ donné par (5) est biaisé.

1.2) La méthode des Moindres Carrés Etendus (MCE)

L'objectif de la méthode est de conserver le caractère quadratique du problème d'optimisation à résoudre tout en conduisant à une estimation sans biais. Dans le cas présent la forme générale (3) s'écrit:

$$A(z^{-1})\, y_t = B(z^{-1})\, u_t + C(z^{-1})\, \varepsilon_t \qquad (8)$$

On a donc posé $P(z^{-1}) = C(z^{-1})$ et $Q(z^{-1}) \equiv A(z^{-1})$

Par référence au formalisme précédent on pose:

$$\underline{\theta}^T = [b_o, ..., b_m, a_1, ..., a_n, c_1, c_2, ..., c_p]$$

$$\underline{z_t}^T = [u_t, ..., u_{t-m}, -y_{t-1}, ..., -y_{t-n}, \varepsilon_{t-1}, ..., \varepsilon_{t-p}] \qquad (9)$$

$$\varepsilon_t = y_t - \underline{z_t}^T.\underline{\theta}$$

$$t = n+1, ..., N$$

Pour que ε_t reste linéaire en $\underline{\theta}$ il faut que les ε_{t-i} soient connus. Pour y parvenir il suffit d'opérer séquentiellement en estimant successivement θ et ε_t suivant un schéma du type Moindres Carrés "en ligne". Celui retenu [Doncarli et al., 1978] est globalement stable, à la différence par exemple de la méthode de Panuska [1984].

1.3) La méthode du maximum de vraisemblance

Cette approche prend pour support la forme (8) en admettant cette fois que c_i et ε_{t-i} sont inconnus et à estimer simultanément. Soit J le critère à minimiser:

$$\theta_{MV} = \arg\min_{\theta} \left(J = \sum_{t=n+1}^{N} \varepsilon_t^2(\theta)\right) \qquad (10)$$

où ε_t et $\underline{\theta}$ sont les paramètres définis en (9).

On notera alors que J est quadratique par rapport aux a_i et b_i et fortement non linéaire par rapport aux c_k. Le calcul de $\underline{\theta}_{MV}$ passe donc par un algorithme de programmation non linéaire.

2) Minimisation d'une erreur de sortie

Cette approche communément appelée Méthode du Modèle (MM) constitue une démarche très générale applicable à n'importe quel type de modèle, en particulier non linéaire.

En se référant au schéma de la figure 1, on pose maintenant $P(z^{-1}) / Q(z^{-1}) \equiv 1$ de sorte que $\varepsilon_t \equiv b_t$ erreur de sortie. La relation (3) devient:

$$A(z^{-1})\, y_t = B(z^{-1})u_t + A(z^{-1})\, \varepsilon_t$$

$$b_t = \varepsilon_t = y_t - \frac{B(z^{-1})}{A(z^{-1})}\, u_t \qquad (11)$$

alors: $$\theta_{MM} = \arg\min_{\theta} \left[J = \sum_{t=n+1}^{N} \varepsilon_t^2(\theta)\right] \qquad (12)$$

avec ε_t donné par (11) et $\underline{\theta}$ par (6).

Là encore J est quadratique en b_i mais fortement non linéaire en a_i, de sorte que le calcul de $\underline{\theta}_{MM}$ passe par l'utilisation d'un algorithme de programmation non linéaire.

IV - Aspects numériques et informatiques

La robustesse, la portabilité, la maintenabilité sont les qualités nécessaires d'un logiciel visant à être utilisé par tous. Elles doivent rester à l'esprit tout au long de sa conception, pour la gestion des données, le choix des algorithmes de calcul, la structuration du code programme et la composition de la documentation.

1) Techniques numériques

Sur le plan numérique, on a essentiellement deux problèmes distincts à résoudre, la minimisation sans contrainte d'une fonction quadratique et celle d'une fonction non linéaire quelconque.

1.1) Problème de Moindres Carrés

On peut l'énoncer de la façon suivante:

$$\underline{x} = \arg\min_{\underline{x}} \|A\underline{x} - b\| \qquad (13)$$

où A est une matrice m x n (m > n) de rang n, $x \in R^m$. On sait qu'une forme explicite de $\underline{x}$ est donnée par :

$$\underline{x} = (A^T A)^{-1} A^T \underline{b} \qquad (14)$$

qui est la forme classique utilisée dans (5). Bien que largement utilisée, cette forme est à bannir car elle dégrade considérablement la précision que l'on peut attendre pour $\underline{x}$.

Soit Q une matrice orthogonale telle que:

$$QA = \left[\frac{U}{O}\right]$$

où U est triangulaire supérieure de dimension n, alors d'après (13) il vient:

$$\|Ax - b\|^2 = \|QAx - Qb\|^2 = \left\|\left[\frac{U}{O}\right]x - \left[\frac{b_1}{b_2}\right]\right\|^2$$

$$= \|Ux - b_1\|^2 + \|b_2\|^2$$

On a noté:

$$\left[\frac{b_1}{b_2}\right] = Qb$$

D'où $\underline{x} = U^{-1}\underline{b}_1$; $\|\underline{b}_2\| = \|A\underline{x} - \underline{b}\|$ (16)

La solution est donc obtenue par simple remontée triangulaire. On notera que U est identique à la factorisation de Choleski de A^TA. La matrice Q peut être obtenue soit par transformation de Householder, soit par rotations de Givens. On a retenu cette dernière solution car elle permet un traitement séquentiel avec un coût mémoire optimum. Son application aux Moindres Carrés Etendus peut être faite facilement [Laporte et al., 1984].

Des informations relatives aux paramètres calculés peuvent être déduites de leur matrice de covariance estimée par:

$$C = \frac{\varepsilon^2_m}{m-n} (U^T U)^{-1} \qquad (17)$$

où m et n sont les dimensions de la matrice A.

1.2) Optimisation non linéaire

Il serait trop long d'entrer ici dans le détail de ces techniques. L'expérience a conduit à choisir une technique du type quasi-Newton que l'on a appliquée aux deux critères (10) et (12). Celle-ci incorpore une technique de mise à jour d'une estimation du Hessien sous forme factorisée qui garantit la stabilité numérique des calculs. L'itération courante est du type :

$$\underline{x}_{t+1} = \underline{x}_t + \lambda_t \underline{d}_t$$

$$\underline{d}_t = (L_t D_t L_t^T)^{-1} \underline{g}_t \qquad (18)$$

où $\underline{g}_t$ est le gradient analytique du critère, λ_t un scalaire obtenu par une technique assurant la convergence. Enfin LDL^T est la factorisation du Hessien estimé avec L triangulaire inférieure à diagonale unité et D diagonale.

2) Environnement interactif de l'identification

Vis-à-vis de l'utilisateur, la communication avec le système informatique est un élément fondamental. Un logiciel performant restera inemployé si son utilisation nécessite un trop gros effort de compréhension. Toute l'attention doit donc être portée sur la définition du dialogue homme-machine au niveau de la tâche à réaliser et de l'acquisition des données correspondantes mais aussi au niveau de l'expression des résultats. L'utilisateur dispose d'un ensemble de mesures du système qu'il étudie. Il a déjà envisagé la structure générale du modèle qui les relie. Son propos est de trouver rapidement les valeurs des paramètres de ce modèle pour ensuite le valider ou le modifier. C'est grâce à un ensemble de résultats qui ne se limitent pas à la seule valeur d'un jeu de paramètres que l'utilisateur pourra juger de la qualité du modèle obtenu et en cas d'échec pourra décider des modifications à y apporter.

La généralité du logiciel impose la conception d'un interface interactif sophistiqué qui doit diriger au plus court l'utilisateur parmi l'ensemble des possibilités offertes en ne lui proposant que les choix propres à son problème. On doit lui permettre également tous les retours en arrière qu'impose l'identification complète d'une structure prédéterminée (changement de l'ordre du système, changement de méthode d'identification...) en évitant les questions redondantes (figure 3). Si l'utilisateur est satisfait de son modèle, il sort du bloc d'identification. Sinon, il a possibilité de "remonter" à tous les niveaux (définition d'un nouveau modèle, choix d'une autre méthode). Un système de dialogue avec réponses par défaut minimise le nombre de renseignements qu'il doit alors redonner.

Nous avons vu précédemment que l'identification ne se limitait pas à un simple calcul de paramètres mais nécessitait l'estimation de la précision de ce calcul et de paramètres secondaires:

- Calculs des écarts-type
- Calcul des gains
- Calcul des pôles et zéros
- Etude de l'écart minimisé.

La comparaison des mesures et de la simulation du modèle identifié ainsi que des entrées apparaît systématiquement sur le terminal. On peut ainsi mieux appréhender les causes des écarts importants.

Grâce à toutes ces données, l'utilisateur peut décider sur le champ des modifications à apporter à son modèle et reprendre une nouvelle identification ou bien étudier les résultats hors ligne. De toute façon, le dernier modèle identifié sera sauvegardé dans la structure de données du système SIRENA à la sortie de la directive Modélisation-Système.

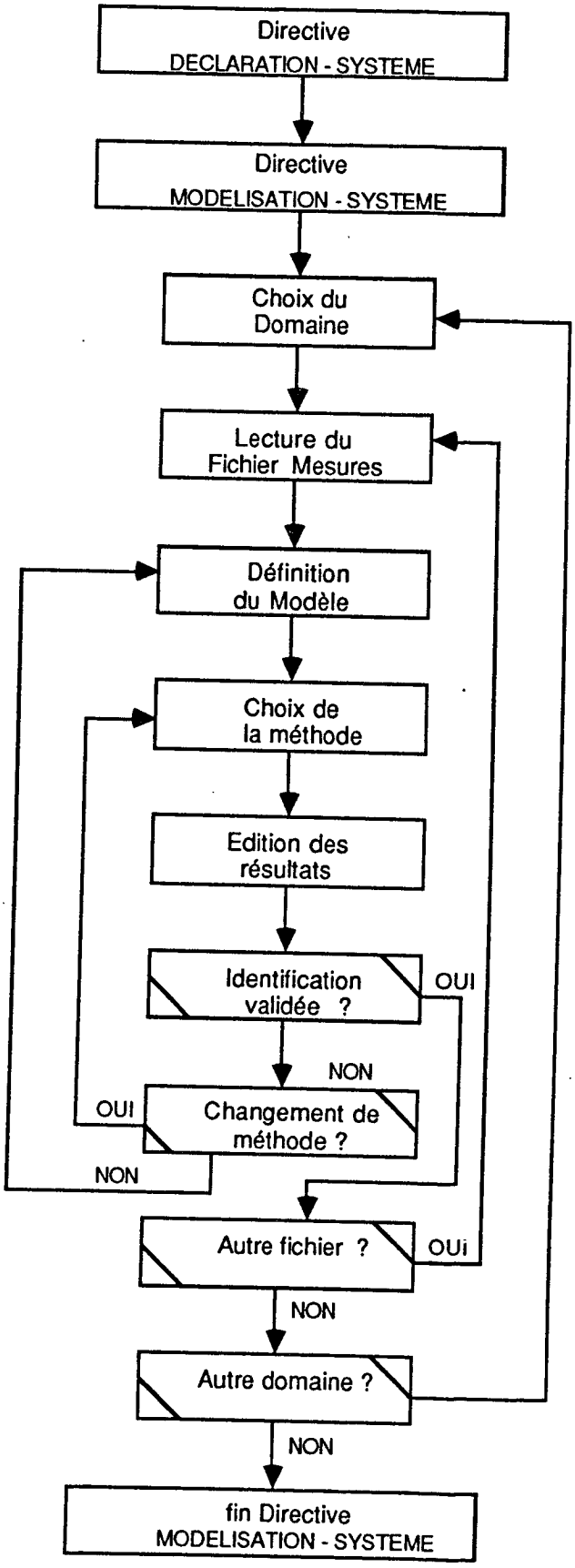


Figure 3 - Dialogue interactif pour une identification

V - Identification récursive

Nous avons déjà signalé plus haut que certaines méthodes, plutôt que de traiter les données en bloc, les traitent séquentiellement. On effectue alors une mise à jour permanente des paramètres du procédé en fonction de chaque mesure. La connaissance des modifications des valeurs des paramètres au cours du temps peut être très intéressante du point de vue analyse du système: soit que ces paramètres évoluent effectivement et qu'il soit intéressant de suivre ces variations: soit qu'on s'attende à les trouver constants et que leur variation permette de détecter un défaut dans les mesures ou un défaut du modèle (dont les paramètres varient pour qu'il soit capable de suivre le comportement du système).

De façon générale, un algorithme d'identification récursive peut être considéré comme un asservissement: la sortie du modèle est asservie à suivre celle du procédé en adaptant ses paramètres (figure 4).

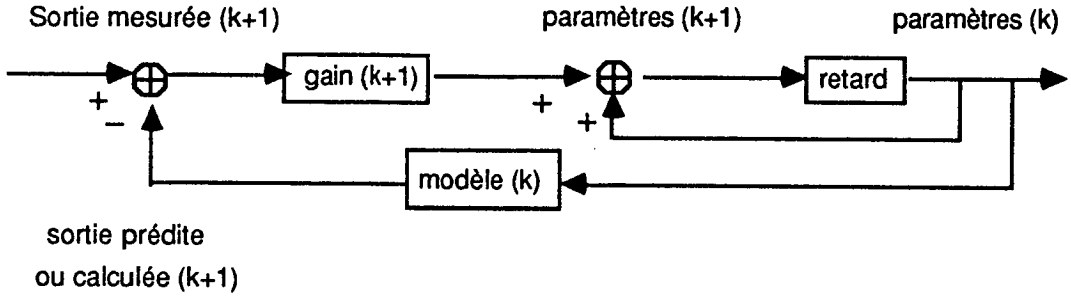


Figure 4 - Principe de l'identification récursive

Comme dans tout asservissement, un gain réglable permet d'obtenir un "bon" comportement du système. C'est ce gain qui fait la différence entre toutes les méthodes récursives. L'algorithme récursif le plus simple (correspondant à la méthode des moindres carrés simples vue plus haut) est identique à celui connu sous le nom de filtre de Kalmann. Nous signalons ce fait car ce filtre commence à être cité par les biologistes, qui ne sont pas toujours conscients de ses limites, et qu'il existe depuis longtemps des solutions plus performantes.

Le laboratoire d'Automatique de Grenoble possède un logiciel d'identification récursive, implémentant six méthodes parmi les plus connues sur des modèles linéaires discrets. Bien que parfaitement interactif, il n'a pas atteint le degré de maturité du logiciel décrit plus haut sur le plan numérique et n'est pas dans une version industrielle.

VI - Exemples d'utilisation

Les exemples décrits ci-dessous ayant pour la plupart fait l'objet de publications, ils ne seront décrits que brièvement et le lecteur pourra se référer à la bibliographie pour des explications détaillées.

Ce que nous voulons souligner en préambule c'est qu'on a beaucoup fait ressortir la spécificité des systèmes biologiques du point de vue de l'identification (voir par exemple [Eykoff, 1985]): peu de connaissances a priori; peu de mesures; variabilité dans le temps et d'un sujet à l'autre; non linéarités; difficultés expérimentales dues aux conditions de mesure ou à des considérations éthiques (dans le domaine biomédical entre autre).

Il serait tout aussi intéressant de souligner leurs similitudes avec les procédés technologiques: entre autre, toute l'approche dite compartimentale est identique à celle de l'automaticien (analyse en fonction des variables internes du système et non des variables externes). Toute expérience faite autour d'un point d'équilibre (état "normal" du sujet, du système) permet d'envisager une linéarisation (en particulier cas des mesures avec traceurs radio-actifs). Alors, si le modèle est identifiable (autrement dit si le problème est correctement posé), il se réduit à une fonction de transfert et toutes les méthodes envisagées ci-dessus s'appliquent . Nous allons dans ce qui suit donner deux exemples plus originaux, dans le domaine de l'environnement.

Le premier exemple se réfère aux fluctuations de la nappe phréatique en bordure du Rhône: l'objectif était de classer les points de mesure, certaines zones apparaissant directement liées au fleuve et d'autres plus dépendantes des précipitations. Sur les neufs zones analysées (10 années de mesure pour chacune), un modèle linéaire discret à quatre paramètres a permis une très bonne description des fluctuations en fonction des trois entrées envisagées (pluie, débit du fleuve, température). On a pu ainsi découper le terrain en zone pluviale et zone fluviale selon que le gain du système était fort par rapport à l'une ou l'autre des entrées. Ce découpage a coïncidé d'une part avec les idées a priori des gens de terrain; d'autre part avec le découpage effectué à l'aide d'une analyse en composantes principales par des statisticiens. De plus, le modèle a montré des capacités prédictives efficaces à la semaine (utilité pour les gens du terrain de prévoir des débordements pour préparer leur matériel) [Gentil et al., 1984].

Le deuxième exemple est beaucoup plus complexe, puisqu'il s'agit d'un écosystème lacustre. Ces systèmes sont généralement analysés à travers des modèles non linéaires comportant un grand nombre de paramètres peu ou pas connus d'avance et "calibrés" en fonction de mesures imprécises et peu nombreuses, ce qui amène évidement à se poser des questions sur la validité scientifique d'une telle approche. L'analyse des données par un modèle linéaire [Gentil, 1984] a pu montrer quelles étaient les entrées les plus influentes; que certains sous-modèles linéaires étaient très performants; que les non-linéarités influaient à certaines périodes de l'année uniquement (ceci étant perçu grâce aux variations des paramètres lors d'une estimation récursive).

VII - Ouverture vers les systèmes experts

Nous avons vu que le problème de l'identification de systèmes linéaires, maintenant bien dominé par les automaticiens (nombreuses méthodes, logiciels spécifiques évitant tout effort de programmation) pourrait aider les bio-écologistes à aborder certains de leurs problèmes. Il nous faut maintenant souligner que malgré cette maturité, identifier un système est loin d'être trivial [Gentil et al., 1985].

Le novice placé devant un logiciel du type décrit ci-dessus n'a pas souvent une stratégie cohérente. Nous avons pu vérifier ceci aisément à travers les réactions des étudiants qui l'utilisent (et qui ont pourtant tous frais en mémoire les concepts exposés dans un cours magistral d'identification). Ils choisissent souvent au hasard, utilisent toujours la même méthode, souvent la plus complexe (se privant ainsi de la souplesse du logiciel). Compte tenu du confort d'utilisation ils empilent un grand nombre de résultats pour différentes structures du modèle; là encore, les valeurs sont choisies au hasard, à moins qu'une stratégie "marteau pilon", ne soit adopté: toutes les combinaisons possibles sont essayées. Il n'est pas toujours facile de faire ensuite un choix parmi tous ces résultats. Le plus gros inconvénient des logiciels robustes est que même si les hypothèses de base des méthodes sont violées (bruit coloré, entrée insuffisamment excitante,...) un résultat numérique est fourni, et de ce fait accepté par le néophyte.

Or il faut analyser soigneusement un résultat pour décider que l'on valide un modèle ou pour reboucler avec une autre structure (degrés de la fonction de transfert différents par exemple) ou avec une autre méthode, ou encore sur un traitement de données (centrage, filtrage,...) ou éventuellement sur une autre expérience (période d'échantillonnage mal choisie,...). Ceci est d'autant plus difficile que tous ces choix sont imbriqués (décider d'utiliser une autre méthode peut dépendre du type d'entrée par exemple). Cela nécessite une bonne connaissance des méthodes elles-même et des propriétés des systèmes de façon plus générale. Cela exige de l'expérience. Cela ne peut être réalisé facilement par de la programmation de type procédural. L'approche système expert parait beaucoup plus prometteuse. Elle permet de séparer clairement la connaissance d'un domaine du mécanisme qui l'exploite.

Toute la connaissance mise en œuvre va tendre vers un but: valider un modèle identifié, parmi d'autres modèles jugés moins performants. Pour cela, des sous-buts seront recherchés, chacun ayant pour objectif de revenir vers le logiciel numérique d'identification pour l'utiliser dans d'autres conditions: changer de méthode et/ou changer la structure du modèle. On peut aussi envisager un autre sous-but: changer les conditions expérimentales, ce qui amène à se servir du logiciel sur d'autres fichiers d'enregistrements [entrées-sorties].

Les analyses de résultats d'identification s'organisent en trois niveaux de connaissances: la critique du modèle que l'on vient d'obtenir, la comparaison de ce modèle avec d'autres modèles correspondant aux mêmes conditions expérimentales et enfin l'examen du contexte expérimental lui-même qui permet de comparer le modèle avec d'autres modèles établis dans d'autres contextes pour valider définitivement ou non l'identification. Il faut remarquer que, bien qu'en principe cette dernière étape soit nécessaire à un travail sérieux, elle n'est pas toujours mise en œuvre de façon rigoureuse; la comparaison pourra alors porter sur une idée plus intuitive que se fait l'utilisateur, du procédé: ordre de grandeur de certaines caractéristiques par exemple.

L'identification des systèmes linéaires nous paraît un excellent exemple pour évaluer la faisabilité des systèmes experts: problème limité, connaissances assez bien définies, nécessité d'un couplage entre algorithmes numériques et utilisation de la connaissance. Ce domaine pourrait servir très efficacement de banc d'essai pour tester différentes stratégies de représentation et d'exploitation des connaissances.

Bibliographie

DONCARLI C., DE LARMINAT Ph., "Analyse de la stabilité globale d'un algorithme d'identification récursive des systèmes linéaires stochastiques discrets". RAIRO Automatique, vol. 12, n° 3, 1978, pp 269-276.

EYKOFF P., "Biomedical identification: overview, problems and prospects". IFAC Symp. on Identification and Parameter Estimation. York (GB), juillet 1985.

FOULARD C., GENTIL S., SANDRAZ J.-P., "Commande par calculateur: de la théorie aux applications". Ed. Eyrolles ,1986.

GENTIL S., "Linear and recursive identification for ecosystems modelling". Ecol. Model., 21, 21-33, 1983-84.

GENTIL S., KOSMELJ K., LACHET B., LAPORTE P., PAUTOU G., "Classification statistique et modélisation des niveaux de la nappe phréatique prés de Bréguier-Cordon en relation avec les apports en eau et la température". Revue de géographie Alpine, janvier 1984.

GENTIL S., RECHENMANN F., "Identification des procédés et Intelligence Artificielle". Congrès "Automatique: des outils pour demain". Toulouse, octobre 1985.

LAPORTE P., GENTIL S., BARRAUD A., "Un logiciel d'identification assité par ordinateur". Colloque SEE, Nice, 1984.

LAWSON L., HANSON C., "Solving least squares problems". Prentice Hall, 1974.

PANUSKA V., "An adaptative least squares identification Method". Proc. 8th. IEEE Symp. Adaptative Process, 1984.

WALTER E., "Identifiability of parametric models" Pergamon Press.

YEM Y., CHOUMILIVONG K., BARRAUD A., "SIRENA: un outil de CAO pour l'automatique". MICAD 84. 27 février - 2 mars 1984, Paris.

ESTIMATION INITIALE DES PARAMETRES D'UN SYSTEME DIFFERENTIEL LINEAIRE EN FONCTION DES PARAMETRES APPLICATION AUX MODELES DE CROISSANCE NON LINEAIRES

François HOULLIER* et **Alain PAVÉ**
Laboratoire de Biométrie, U.A. CNRS 243 "Biologie des Populations"
Université Claude Bernard - Lyon 1
43, Bld du 11 Novembre - 69622 Villeurbanne Cedex

* Inventaire Forestier National
Cellule Evaluation de la Ressource
Place des Arcades - 34970 Maurin/Lattes

Introduction

L'estimation des paramètres de modèles non linéaires est généralement réalisée au moyen d'algorithmes numériques itératifs qui requièrent l'existence d'estimations initiales préalables et l'évaluation des fonctions de sensibilité du modèle (Bard, 1974).

Notre but est de présenter une méthode qui permette d'obtenir ces estimations initiales. Deux cas sont considérés:

- Celui des systèmes différentiels linéaires en fonction des paramètres: la méthode consiste à intégrer numériquement le système différentiel en conservant sa structure linéaire, puis à procéder à une régression linéaire multiple. On traite d'abord du cas d'une seule équation différentielle (§ 1). Pour ce qui concerne l'estimation finale des paramètres, on présente ensuite une technique (classique) d'évaluation des fonctions de sensibilité (§ 2). La généralisation de l'estimation initiale au cas de systèmes différentiels à plusieurs variables est discutée au § 5.

- Celui des courbes de croissance univariées analytiques dont la dérivée se présente sous la forme d'une équation différentielle ordinaire linéaire (en fonction des paramètres): la linéarité de cette équation permet de se ramener au cas précédent (soit directement, soit indirectement par une reparamétrisation qui linéarise l'équation; cf. § 3). Des exemples concrets sont proposés: ils concernent quatre modèles "classiques" de la dynamique des populations (cf. § 4).

1 - Estimation initiale des paramètres d'un modèle différentiel, linéaire en fonction des paramètres

On considère un modèle déterministe d'une variable x, qui se présente sous la forme d'une équation différentielle ordinaire, $x' = \phi(x, t, \theta_1, \ldots, \theta_p)$, dont on suppose qu'elle est linéaire en fonction des paramètres θ_j $(1 \leq j \leq p)$

$$\frac{dx}{dt} = \sum_{j=1}^{p} \theta_j f_j (x, t) \qquad [1]$$

avec $x(t_0) = x_0 = \theta_0$ et où $f_j(x, t)$, $1 \leq j \leq p$, sont des fonctions de x et t *a priori* connues;

x_0 est la condition initiale et constitue le $p+1^{ème}$ paramètre du modèle intégré.

On cherche à estimer les paramètres de ce modèle, θ_j $(0 \leq j \leq p)$, à partir d'une série chronologique univariée $((t_k, x(t_k)), 0 \leq k \leq n-1, t_0 \leq t_1 \leq \ldots \leq t_k \leq \ldots \leq t_n)$, où n est le nombre de dates de mesure, t_k la $k^{ème}$ date de mesure et $x(t_k)$ la valeur de x mesurée à t_k.

Remarquons que l'on ne s'intéresse pas aux seuls paramètres de l'équation différentielle, mais à ceux du modèle différentiel formé de ladite équation complétée par une condition initiale. L'identification porte donc sur la variable x et non pas sur sa dérivée par rapport au temps. Notons alors que x ne s'exprime pas nécessairement de façon explicite en fonction du temps. On se place dans le cas où les valeurs de θ_j et l'intervalle $[t_0, t_k]$ sont tels qu'il y a existence et unicité de la solution du modèle différentiel.

L'intégration terme à terme de l'équation différentielle permet d'obtenir:

$$x(t) = \theta_0 + \sum_{j=1}^{p} \theta_j \int_{t_0}^{t} f_j (x, \tau)\, d\tau \qquad [2]$$

Pour estimer les paramètres du modèle dynamique, on doit de plus spécifier un modèle d'erreur. On suppose que:

$$x(t) = F(t, \theta_0, \theta_1, \ldots, \theta_p) = \theta_0 + \sum_{j=1}^{p} \theta_j \int_{t_0}^{t} f_j (x, \tau)\, d\tau + e(t) \qquad [3]$$

et que "l'erreur" aléatoire, e(t), vérifie les hypothèses classiques:

$$\begin{aligned} &E[e(t)] = 0 \\ &Var[e(t)] = \sigma^2 \\ &Cov[e(t),e(t')] = \delta_{(t-t')}\, \sigma^2 \quad (\text{avec } \delta(t-t') = 1 \text{ si } t = t' \text{ et } \delta(t-t') = 0 \text{ sinon}) \end{aligned} \qquad [4]$$

L'équation [3] appliquée aux dates de mesure t_k donne un système linéaire en fonction des paramètresθ_j $(0 \leq j \leq p)$:

$$x(t_k) = \theta_0 + \sum_{j=1}^{p} \theta_j \int_{t_0}^{t_k} f_j (x, \tau)\, d\tau + e(t) \;;\; 1 \leq k \leq n \qquad [5]$$

Sous ces hypothèses, la théorie usuelle de la régression linéaire multiple s'applique au système [3] muni de [4] pour estimer les paramètres θ_j $(0 \leq j \leq p)$: on utilise la méthode des moindres carrés ordinaires. Si le modèle d'erreur [4] est trop restrictif, par exemple si on suppose que les erreurs admettent une autocorrélation temporelle, on doit plutôt utiliser la méthode des moindres carrés pondérés, voire celle du maximum de vraisemblance (**Beck et Arnold**, 1977; **Méssean**, même volume).

Du point de vue numérique, les intégrales peuvent être évaluées à partir des données expérimentales $x_i(t_k)$ en utilisant les méthodes classiques - méthode des trapèzes, interpolation parabolique ou par des fonctions splines d'ordre 3 - (**Pavé**, 1980); par exemple, si on adopte la méthode des trapèzes:

$$\int_{t_0}^{t_k} f_j\,(x(\tau),\tau)\,d\tau \quad \text{est estimée par} \quad \frac{1}{2}\sum_{h=0}^{k-1}[f_j(x(t_{h+1}),t_{h+1}) + f_j(x(t_h),t_h)]\,[t_{h+1}-t_h]$$

Les estimations ainsi obtenues ne peuvent être considérées que comme des ***estimations initiales***. En effet, le modèle complet [3] n'est valide que si les valeurs prédites de x sont reportées dans les intégrales. Ces valeurs étant inconnues, on utilise en fait les valeurs mesurées. Il s'ensuit un cumul des erreurs qui est incompatible avec le modèle d'erreur défini précédemment (**Pavé**, 1980, page 65).

2 - Estimation finale des paramètres d'un modèle différentiel

On considère la situation définie précédemment, qui se compose:

. d'une part d'une relation déterministe (ici, une équation différentielle ordinaire et une condition initiale);

. d'autre part, d'un modèle d'erreur qui détermine un ***critère*** dont la valeur est extrémale pour les valeurs estimées des paramètres (par exemple: somme des carrés des écarts ou fonction de vraisemblance, (**Beck et Arnold**, 1977).

Les diverses méthodes d'estimation finale des paramètres se présentent alors sous forme d'une ***méthode de minimisation*** d'un critère. Dans le cas où le modèle intégré n'est pas linéaire en fonction des paramètres, il s'agit de méthodes ***itératives*** qui requièrent (**Bard**, 1974):
. des ***estimations initiales*** des paramètres,
. la valeur du ***gradient du critère*** à chaque itération (certaines méthodes nécessitent de plus la donnée du Hessien de ce critère).

Dans le cas d'un modèle différentiel linéaire en fonction des paramètres, les estimations initiales sont fournies par la méthode présentée au § 1. Leur qualité conditionne largement le succès et la rapidité de la convergence de l'algorithme vers les valeurs qui minimisent effectivement le critère. D'autres possibilités existent cependant:

- linéarisation directe du modèle intégré lorsque c'est possible. Cependant, dans certains cas, les performances sont moins bonnes que pour l'estimation sur l'équation différentielle linéaire, ou linéarisée (cf., par exemple, le cas de la linéarisation du modèle exponentiel par transformation logarithmique qui peut être moins performante, au sens de la proximité du résultat final, que l'estimation sur l'expression différentielle (**Pavé**, 1982));

- recherche visuelle d'un bon ajustement; cette voie a été en particulier développée par **Rousseau** (**Rousseau** et al, 1986) dans le cas de certaines courbes de croissance univariées possédant des paramètres phénoménologiques (asymptotes, point d'inflexion) dont la représentation graphique est aisée.

La valeur du gradient du critère s'obtient à partir des fonctions de sensibilité, $g_j = (\partial x/\partial\theta_j)$. Quand il n'existe pas de forme analytique du modèle, ces fonctions ne sont pas explicitement connues. Pour les calculer, on considère:

$$\frac{d}{dt}\left(\frac{\partial x}{\partial \theta_j}\right) = \frac{\partial \phi}{\partial \theta_j} + \frac{\partial \phi}{\partial x}\frac{\partial x}{\partial \theta_j} \qquad (1 \le j \le p)$$

soit $$\frac{dg_j}{dt} = f_j + g_j \sum_{k=1}^{p} \theta_k \frac{\partial f_k}{\partial x} \quad \text{où} \quad g_j = \frac{\partial x}{\partial \theta_j} \qquad (1 \le j \le p) \qquad [6]$$

on a de plus $$\frac{d}{dt}\left(\frac{\partial x}{\partial \theta_0}\right) = \frac{\partial \phi}{\partial x}\frac{\partial x}{\partial \theta_0}$$

soit $$\frac{dg_0}{dt} = g_0 \sum_{k=1}^{p} \theta_k \frac{\partial f_k}{\partial x} \quad \text{où} \quad g_0 = \frac{\partial x}{\partial \theta_0} \qquad [7]$$

et les conditions initiales $$\left(\frac{\partial x}{\partial \theta_j}\right)_{t=t_0} = 0 \text{ si } j \neq 0 \text{ et } \left(\frac{\partial x}{\partial \theta_0}\right)_{t=t_0} = 1 \qquad [8]$$

Les paramètres θ_j $(0 \le j \le p)$ étant fixés, il s'agit donc, pour évaluer les fonctions de sensibilité à ces paramètres, de résoudre numériquement un système différentiel en g_j, formé des équations [6], [7] et [8] (cf., par exemple **Vila**, 1982).

3 - Estimation initiale des paramètres d'une courbe de croissance analytique non linéaire

On considère un modèle univarié s'exprimant sous une forme analytique et non linéaire en fonction des paramètres:

$$x(t) = F(t, \theta_0, \theta_1, \ldots, \theta_p) + e(t)$$

où θ_0 représente une condition initiale ($\theta_0 = x(t_0)$) et e(t) une erreur aléatoire.

Pour estimer les paramètres $\theta_0, \theta_1, \ldots, \theta_p$ du modèle, on dispose comme précédemment d'une série chronologique $((t_k, x(t_k)), 0 \le k \le n-1)$.

On suppose que le modèle est différentiable (en fonction du temps) et que sa forme différentielle s'exprime de façon linéaire en fonction des paramètres θ_j $(1 \le j \le p)$ ou d'autres paramètres, notés β_j $(1 \le j \le r)$, obtenus par reparamétrisation à partir des θ_j. On est ainsi ramené au cas précédent

$$\frac{dx}{dt} = \sum_{j=1}^{r} \beta_j f_j(x, t) \quad \text{où} \quad \beta_j = \psi_j(\theta_1, \ldots, \theta_p) \quad \text{pour } 1 \le j \le r$$

$$x(t) = \beta_0 + \sum_{j=1}^{r} \beta_j \int_{t_0}^{t} f_j(x, \tau)\, d\tau + e(t) \quad \text{où} \quad \theta_0 = \beta_0 = x(t_0)$$

L'estimation initiale des paramètres β_j ($0 \leq j \leq r$) est obtenue par régression linéaire. Les paramètres θ_j sont alors estimés en résolvant le système (statique):

$$\beta_j = \psi_j (\theta_1, \ldots, \theta_p)$$
$$\beta_0 = \theta_0$$

Une condition nécessaire (mais non suffisante) pour que ce système admette une solution unique est que r = p, où r+1 est le nombre de nouveaux paramètres, β_j, et p+1 le nombre de paramètres initiaux, θ_j :
. si r < p, le modèle n'est pas identifiable par cette méthode (on ne peut pas estimer individuellement tous les paramètres θ_j);
. si r > p, le modèle est suridentifié: on dispose ainsi de plusieurs estimations, non nécessairement identiques, pour l'ensemble des paramètres θ_j.

Pour certains modèles analytiques, il ne s'avère pas possible de linéariser l'équation différentielle associée de telle sorte que la correspondance entre les paramètres initiaux et les paramètres du modèle linéarisé soit biunivoque (cf modèle de Johnson-Schumacher §4.3): il n'est alors pas possible d'appliquer la méthode.

4 - Exemples

Cette méthode d'estimation initiale a été utilisée avec succès dans de nombreux cas: modèle logistique, modèle de Gompertz [**Pavé** et al, 1986], et même le modèle exponentiel pour lequel cette méthode est souvent préférable à la linéarisation classique par transformation logarithmique des données [**Pavé**, op. cit.]. Ici nous ne reprenons que quelques exemples à des fins d'illustration, notamment le modèle de Kostitzin qui fut le premier sur lequel cette méthode a été employée [**Pavé**, 1979]. Enfin, on pourra se reporter à la contribution de Pavé "interprétation et construction de modèles de la dynamique des populations à l'aide de schémas fonctionnels" (même volume) pour trouver des informations et des données expérimentales complémentaires.

4.1 - Modèle de Monod (Corman et Pavé, 1983)

Son expression différentielle est

$$\frac{dx}{dt} = \frac{\theta_1 (\theta_2 - x)\, x}{\theta_2 + \theta_3 - x}$$

Cette équation n'est pas linéaire en fonction de θ_2 et de θ_3, cependant en multipliant la partie droite par le dénominateur de la partie gauche et en intégrant, il vient :

$$x^2 = (x_0)^2 + \beta_1 (x - x_0) + \beta_2 \int_{t_0}^{t} x \, d\tau + \beta_3 \int_{t_0}^{t} x^2 \, d\tau$$

avec $\beta_1 = 2(\theta_3 + \theta_2)$, $\beta_2 = -2\,\theta_1\theta_2$ et $\beta_3 = 2\,\theta_1$. En outre, on peut retrouver facilement les paramètres classiques du modèle quand il est utilisé pour décrire la croissance d'une population bactérienne dans un milieu limité en substrat : $\theta_1 = \mu_0$ (taux de croissance maximum), $\theta_2 = R\, s_0 + x_0$ et $\theta_3 = R\, K_S$, s_0 est la quantité initiale en

substrat, R est le rendement de la croissance (quantité de substrat nécessaire à la croissance d'un unité de biomasse) K_S peut être interprétée comme l'inverse d'une constante d'affinité de la bactérie pour le substrat.

Par exemple, dans l'article de **Corman et Pavé** (op. cit), on trouve des comparaisons entre les estimations initiales et finales des paramètres de ce modèle dans trois cas, et également un tableau comparant valeurs expérimentales, valeurs calculées avec les estimations initiales et valeurs calculées avec les estimations finales. On ne reprendra ici que les comparaisons entre estimations initiales et finales (tableau 1).

	Escherichia Coli sur glucose		*Nitrobacter win.* sur NO_2Na		*Lactobacillus delb.*	
	(e.i.)	(e.f.)	(e.i.)	(e.f.)	(e.i.)	(e.f.)
μ_0	0.886	0.886	0.0382	0.0387	0.680	0.692
K_S	1.54	1.51	6.0	7.5	52.7	57.0
R	0.307	0.309	0.14×10^{-8}	0.13×10^{-8}	0.0948	0.0960

Tableau 1 - Modèle de Monod : comparaison des estimations initiales (e.i.) et finales (e.f.) des paramètres, pour trois ensembles de données sur la croissance de trois populations bactériennes en milieu limité, d'après (**Corman et Pavé**, 1983).

4.2 - Modèle logistique généralisé

Sous forme analytique, le modèle logistique généralisé ou modèle de Nelder (suivant la paramétrisation proposée par **Debouche**, 1979) s'écrit:

$$x = \theta_1 \left[1 + \theta_4 \, e^{\frac{\theta_2 - t}{\theta_3}} \right]^{-\frac{1}{\theta_4}}$$

Sous forme différentielle, on a:

$$\frac{dx}{dt} = \frac{x}{\theta_3 \theta_4} \left[1 - \left(\frac{x}{\theta_1} \right)^{\theta_4} \right]$$

soit p=3. Cette équation n'est pas linéaire en fonction de θ_4 et θ_1. On considère alors deux cas, selon que le paramètre θ_4 est ou non a priori fixé.

4.2.1 - ***Si*** θ_4 ***est a priori fixé***, on obtient:

$$\frac{dx}{dt} = \beta_1 x - \beta_2 x^{1+\theta_4} \quad \text{avec} \quad \beta_1 = \frac{1}{\theta_3 \theta_4} \quad \text{et} \quad \beta_2 = \frac{\theta_1^{-\theta_4}}{\theta_3 \theta_4}$$

La méthode décrite au paragraphe 3 s'applique donc:

. de β_1, on déduit θ_3;

. de β_2, on déduit θ_1;

. on considère la relation supplémentaire:

$$x = x(t_0) = \theta_1 \left[1 + \theta_4 \, e^{\frac{\theta_2 - t_0}{\theta_3}} \right]^{-\frac{1}{\theta_4}}$$

pour déduire θ_2.

4.2.2 - *Si* θ_4 *n'est pas fixé* et si $\theta_4 \approx 0$, on cherche à estimer ce dernier paramètre en linéarisant l'équation différentielle correspondant au modèle, au voisinage de $\theta_4 = 0$. En développant x^{θ_4} au second ordre, il vient

$$x^{\theta_4} = e^{\theta_4 \ln x} \approx 1 + \theta_4 \ln x + \theta_4^2 \frac{[\ln x]^2}{2}$$

L'équation différentielle approchée s'écrit alors:

$$\frac{dx}{dt} \approx \beta_1 x + \beta_2 x \ln x + \beta_3 x (\ln x)^2$$

alors

$$x \approx x_0 + \beta_1 \int_{t_0}^{t} x \, d\tau + \beta_2 \int_{t_0}^{t} x \ln x \, d\tau + \beta_3 \int_{t_0}^{t} x (\ln x)^2 \, d\tau$$

où $\beta_1 = \frac{1}{\theta_3 \theta_4} \left(1 - \frac{1}{(\theta_1)^{\theta_4}} \right)$; $\beta_2 = -\frac{1}{\theta_3 \theta_1^{\theta_4}}$;

$$\beta_3 = -\frac{\theta_4}{2\theta_3 \theta_1^{\theta_4}} = \beta_2 \frac{\theta_4}{2}$$

soit $\theta_4 = 2 \frac{\beta_3}{\beta_2}$; $\theta_1 = \left(1 - \beta_1 \frac{\theta_4}{\beta_2} \right)^{1/\theta_4}$; $\theta_3 = -\frac{1}{\beta_2 \theta_1^{\theta_4}}$

et enfin $\theta_2 = t_0 + \theta_3 \ln \left[\frac{1}{\theta_4} \left(\left(\frac{\theta_1}{x_0} \right)^{\theta_4} - 1 \right) \right]$

Exemple: croissance du pin noir (Houllier, 1986)

Les données expérimentales proviennent de l'INRA (Avignon) et portent sur la croissance d'un résineux : le pin noir. Ces données ont été obtenues à partir d'une technique dite "d'analyse de tige".

âge (en années)	11	14	18	22	25	29	32	35	38	42	45	49	52	56	60	64	68	73	78	83	88	95	103
hauteur (en m)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	22.8

L'estimation initiale des paramètres a été obtenue à partir de la méthode décrite ci-dessus.

Dans la pratique, il s'est avéré intéressant de réaliser l'estimation initiale en deux phases:

. on a commencé par estimer les 4 paramètres en utilisant la méthode présentée ci-dessus;

. on fixé ensuite la valeur θ_4 obtenue et on réestime les trois premiers paramètres, θ_1, θ_2 et θ_3 selon la première méthode.

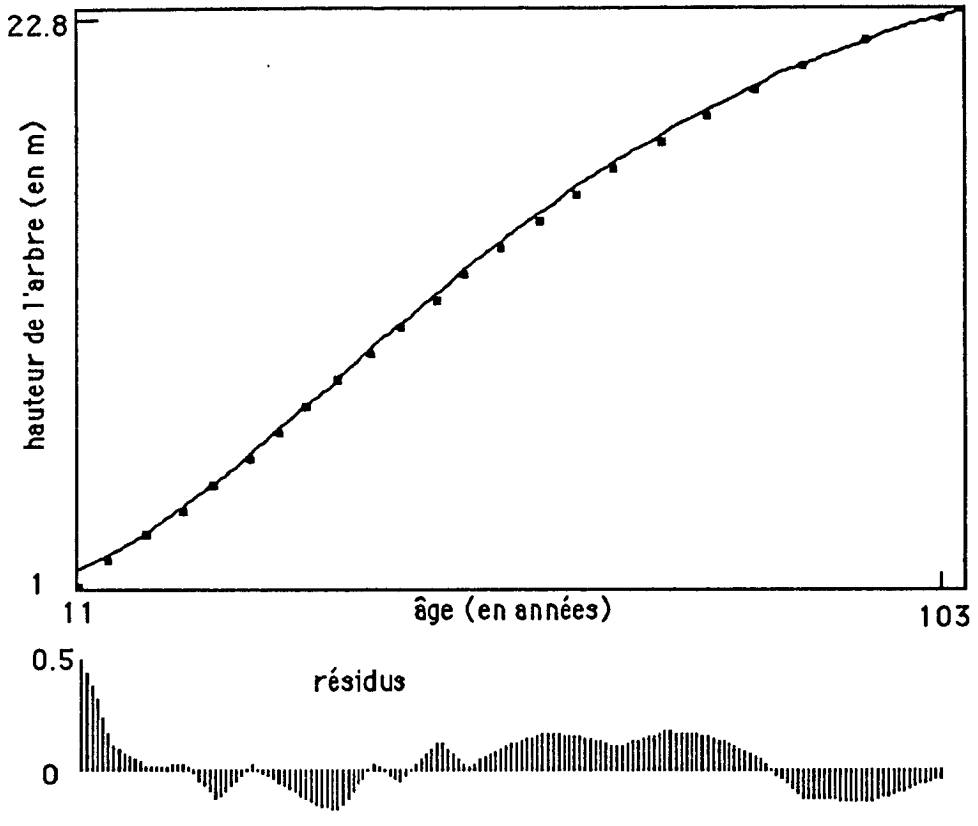


Figure 1 - Estimations initiales des quatres paramètres et résidus (valeurs cal.- valeurs obs.): $\theta_1 = 26.88$; $\theta_2 = 37.41$; $\theta_3 = 36.08$; $\theta_4 = -0.234$. La somme des carrés des écarts est SCE init. = 0.659. Le nombre des séquences des résidus est de 8 (nombre de changements de signes + 1).

Cette façon de procéder, purement empirique, a permis d'améliorer sensiblement les estimations initiales (en termes de sommes des carrés des écarts résiduels), mais elle n'assure pas que la convergence sera plus rapide : celle-ci dépend, en effet, essentiellement de la forme du critère et de la validité de l'approximation linéaire au voisinage du minimum.

L'estimation finale a été réalisée en utilisant une méthode du type Gauss-Marquardt proposée par Fletcher et disponible dans la Bibliothèque d'Harwell.

On remarque que la première méthode donne de bons résultats si on se limite à une comparaison "à vue" des courbes calculées avec les points expérimentaux; le gain obtenu par l'utilisation de la méthode de Gauss-Marquardt est néanmoins sensible lorsqu'on examine les sommes des carrés des écarts (divisée par 3), et la distribution des résidus (le nombre des séquences est de 8 dans le premier cas, et de 15 dans le second).

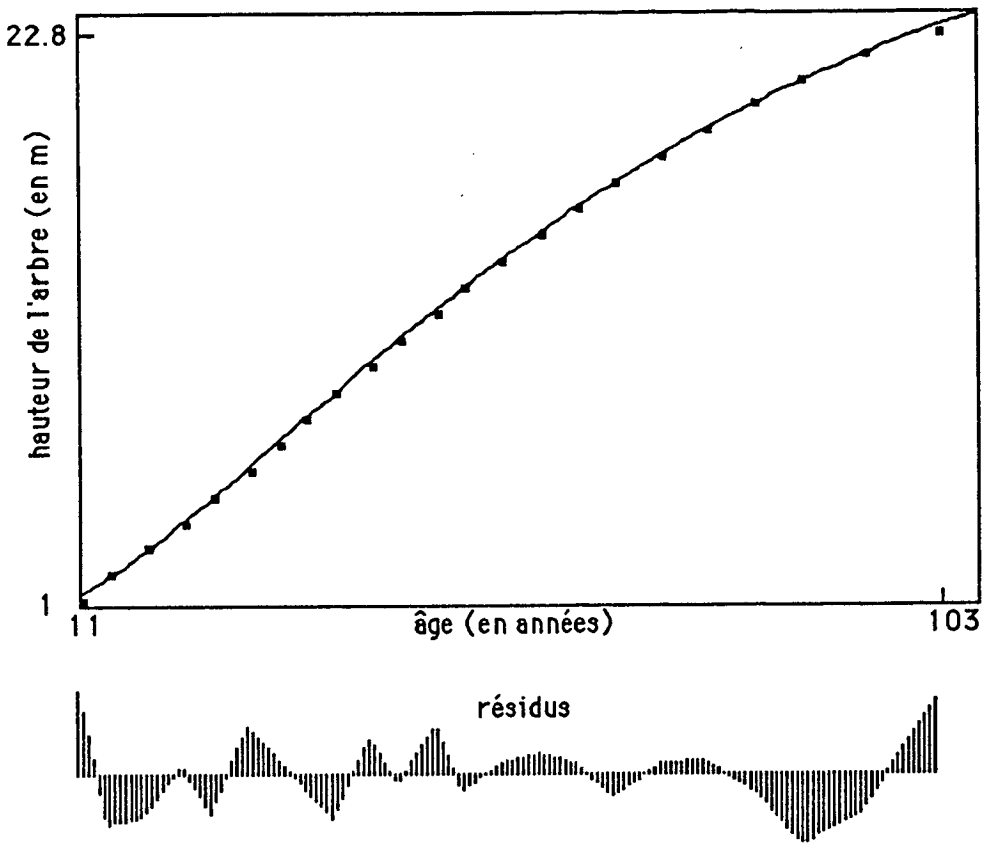


Figure 2 - Estimations finales et résidus (valeurs cal. - valeurs obs.) $\theta_1 = 29{,}63$; $\theta_2 = 32.36$; $\theta_3 = 49.10$; $\theta_4 = -0.528$. La somme des carrés des écarts est SCE fin. = 0.226. Le nombre des séquences des résidus est de 15

4.3 - Modèle de Johnson-Schumacher

Sous forme analytique, le modèle de Johnson-Schumacher (**Debouche**, 1979) est un modèle à trois paramètres qui s'écrit:

$$x(t) = \theta_1 \, e^{-\frac{\theta_3}{\theta_2 - t}}$$

Sous forme différentielle, on a:

$$\frac{dx}{dt} = \frac{[\ln \theta_1]^2}{\theta_3} x - 2 \frac{\ln \theta_1}{\theta_3} x \ln x + \frac{1}{\theta_3} x \, [\ln x]^2$$

La forme intégrée est alors:

$$x(t) = x_0 + \frac{[\ln \theta_1]^2}{\theta_3} \int_{t_0}^{t} x \, d\tau - 2 \frac{\ln \theta_1}{\theta_3} \int_{t_0}^{t} x \ln x \, d\tau + \frac{1}{\theta_3} \int_{t_0}^{t} x \, [\ln x]^2 \, d\tau$$

L'utilisation stricte de la méthode précédente conduit ainsi à une surdétermination des paramètres, puisqueles deux paramètres θ_1 et θ_3 sont liés aux trois paramètres β_1 , β_2 et β_3 où:

$$\beta_1 = \frac{[\ln \theta_1]^2}{\theta_3} \; ; \; \beta_2 = -2\frac{\ln \theta_1}{\theta_3} \text{ et } \beta_3 = \frac{1}{\theta_3}$$

Nous avons alors choisi d'appliquer la méthode et de ne conserver que les relations concernant β_1 et β_3 pour reconstituer les paramètres θ_1 et θ_3 ; enfin θ_2 peut être obtenu grâce à la relation supplémentaire:

$$x(t_0) = \theta_1 e^{-\frac{\theta_3}{\theta_2 - t_0}} \qquad \text{soit} \qquad \theta_2 = t_0 - \frac{\theta 3}{\ln\left(\frac{x(t_0)}{\theta_1}\right)}$$

Tout ceci n'est bien sûr possible que parce qu'il s'agit d'estimations initiales qui sont destinées à être améliorées.

4.4 - Modèle de Kostitzin (Pavé, 1979)

Le modèle de Kostitzin est un modèle intégro-différentiel dont l'équation est:

$$\frac{dx}{dt} = \frac{x}{\theta_3}\left[1 - \frac{x}{\theta_1} + \theta_4 \int_{t_0}^{t} x(\tau)\, d\tau\right] \qquad \text{avec} \qquad x(t_0) = x_0$$

Il se présente donc sous la forme d'une équation linéaire en fonction des paramètres et la théorie décrite au paragraphe 3 s'applique. L'intégrale peut en effet être évaluée numériquement à partir des valeurs expérimentales observées. On trouve un exemple d'utilisation de ce modèle, avec des données expérimentales, dans la contribution de A. Pavé (même volume).

5 - Extension au cas d'un système différentiel linéaire en fonction des paramètres

On considère à présent un modèle déterministe qui se présente sous la forme d'un système différentiel, à q variables x_i, linéaire en fonction des paramètres θ_{ij} ($1 \le i \le q$ et $1 \le j \le p_i$):

$$\frac{dx_i}{dt} = \sum_{j=1}^{p_i} \theta_{ij} f_{ij}(x_1, \ldots, x_q, t) \quad \text{pour} \quad 1 \le i \le q \qquad [9]$$

Remarquons que l'on peut avoir des contraintes liant les paramètres, par exemple $\theta_{ij} = \theta_{i'j'}$. Pour estimer les paramètres de ce système, on dispose d'une série chronologique multivariée $((t_k, (x_i(t_k), 1 \le i \le q)), 1 \le k \le n)$, où k est le nombre de dates de mesure.

L'intégration du système permet d'obtenir:

$$x_i(t) = \theta_{i_0} + \sum_{j=1}^{p_i} \theta_{ij} \int_{t_0}^{t} f_{ij}(x_1, \ldots, x_q, \tau)\, d\tau \qquad (1 \le i \le q) \quad \text{où} \quad \theta_{i_0} = x_i(t_0)$$

On considère ensuite le modèle d'erreur:

$$x_i(t) = \theta_{i_0} + \sum_{j=1}^{p_i} \theta_{ij} \int_{t_0}^{t} f_{ij}(x_1, \ldots, x_q, \tau)\, d\tau + e_i(t) \qquad (1 \le i \le q)\ .$$

On suppose que les erreurs aléatoires, $e_i(t_k)$, vérifient les hypothèses classiques:

$$\begin{aligned} &E[e_i(t_k)] = 0 \\ &\text{Var}[e_i(t_k)] = \sigma_i^2 \qquad (1 \le i \le q \quad \text{et} \quad 1 \le k \le n) \\ &\text{Cov}[e_i(t_k), e_j(t_h)] = \delta_{h,k}\ \delta_{i,j}\ \sigma_i^2 \end{aligned} \qquad [10]$$

L'application aux données fournit le système:

$$x_i(t_k) = \theta_{i_0} + \sum_{j=1}^{p_i} \theta_{ij} \int_{t_0}^{t_k} f_{ij}(x_1, \ldots, x_q, \tau)\, d\tau + e_i(t_k) \quad (1 \le i \le q \text{ et } 1 \le k \le n) \quad [11]$$

Sous ces hypothèses, et si les paramètres θ_{ij} sont tous distincts, la théorie classique de la régression linéaire multiple s'applique au système [11] muni de [10] et on estime indépendamment les paramètres θ_{ij} $(0 \le j \le p_i)$ pour chaque variable i.

Si on suppose au contraire que les erreurs admettent des corrélations croisées, ou si certains paramètres affectent plusieurs variables (ce qui est le cas général), on doit estimer simultanément l'ensemble des paramètres θ_{ij} $(1 \le i \le q,\ 0 \le j \le p_i)$ par la méthode des moindres carrés pondérés. On forme alors un critère qui est une double somme pondérée (sur i et sur k) des écarts entre $x_i(t_k)$ et sa valeur prédite. Le choix de la pondération sur les variables est alors particulièrement délicat puisqu'il détermine l'importance attachée à la bonne reconstitution relative de chacune d'entre elles. Contrairement au choix de la pondération sur les dates de mesure qui est liée à la connaissance du modèle d'erreur, la pondération sur les variables reste totalement subjective et traduit les objectifs de l'expérimentateur.

6 - Conclusion

La méthode proposée au paragraphe 1 et illustrée au paragraphe 4 permet d'obtenir des estimations initiales des paramètres d'un modèle, qui peuvent être ensuite améliorées par l'utilisation d'algorithmes classiques du type Gauss-Marquardt. Ces estimations sont assez bonnes même pour des modèles dont l'identification est délicate (par exemple, le modèle de Monod ou le modèle logistique généralisé). En tout état de cause, nous proposons une solution à un problème généralement mal résolu. De plus les nombreux essais qui ont été faits ont permis d'acquérir une expertise dans l'utilisation de cette méthode qui sera de ce fait disponible pour l'utilisateur dans le logiciel Edora (signalons qu'elle est déjà utilisée dans le logiciel CROISSANCE [cf. la contribution de **B. Rousseau** et **F. Rechenmann** dans le même volume]).

Bibliographie

Bard Y. (1974). Non Linear Parameter Estimation. Acad. Press, New York.

Beck J.V., Arnold K.J. (1977). Parameter Estimation in Engineering and Science. J.Wiley & Sons, New-York, 1977.

Corman A., Pavé A. (1983). On parameter estimation of Monod's bacterial growth model from batch culture data. J. Gen. Appl. Microbio., 29, 91-101.

Pavé A., Corman A., Bobillier-Monot B. (1986) Utilisation et interprétation du modèle de Gompertz, application à l'étude de la croissance de jeunes rats musqués (Ondatra zibethica L.). Biom. Praxim., 26, 123-140.

Debouche C. (1979). Présentation coordonnée de différents modèles de croissance. Revue de Statistique Appliquée, 27, 4, 5-22.

Fresen J.L., Juritz J.M. (1986). A Note on Foss's Method on Obtaining Initial Estimates for Exponential Curve Fitting by Numerical Integration. Biometrics, 42, 821-827.

Houllier F. (1986). Echantillonnage et modélisation de la dynamique des peuplements forestiers: application au cas de l'Inventaire Forestier National. Thèse de doctorat, Université de Lyon, 268p.

Pavé A. (1979). Dynamics of macromolecular populations : a mathematical model of the quantitative change of RNA in the silkgland during the last larval instar. Biochimie, 61, 263-273.

Pavé A. (1980). Contribution à la théorie et à la pratique des modèles mathématiques pour l'analyse dynamique de systèmes biologiques. Etude de quelques cas typiques en biologie cellulaire et moléculaire. Thèse d'état, U.C.B. Lyon.

Pavé A. (1982). Les modèles à compartiments linéaires. In "Modèles Dynamiques Déterministes en Biologie", ed. Lebreton J.D. & Millier C., Masson, Paris.

Pavé A. Interprétation et construction de modèles de la dynamique des populations à l'aide de schémas fonctionnels. [même volume].

Rousseau B., Rechenmann F. Edora: Vers un poste de travail informatique pour l'aide à la modélisation des systèmes dynamiques en biologie.[même volume].

Rousseau B., Pavé A., Rechenmann F., Landau M. (1986). Edora project: Artificial Intelligence approach and Work Station concept to Aid Dynamic modelling in Biology and Ecology. Suppl. Proceed. of the Summer Comp. Simul. Conf. Reno, Nevada, 14-20.

Vila J.P. (1982). Méthodes d'identification des modèles dynamiques. In "Modèles Dynamiques Déterministes en Biologie", ed. Lebreton J.D. & Millier C., Masson, Paris.

LE PROJET EDORA

VERS UN POSTE DE TRAVAIL INFORMATIQUE POUR L'AIDE À LA MODÉLISATION DES SYSTEMES DYNAMIQUES EN BIOLOGIE

Bertrand ROUSSEAU
Laboratoire de Biométrie
UA CNRS 243, Université LYON 1
69622 Villeurbanne Cedex

François RECHENMANN
Laboratoire ARTEMIS/Imag (INRIA)
BP 68
38402 St Martin d'Hères Cedex

Les expérimentateurs sont souvent désarmés face à la diversité des compétences nécessaires pour la modélisation des phénomènes biologiques: informatique, calcul numérique, mathématiques, biométrie, domaine biologique...

Le Projet EDORA vise à fournir aux biologistes un système informatique capable de le guider à toutes les étapes du processus de modélisation: choix ou construction d'un modèle, analyse qualitative, simulation et validation...
Un tel logiciel repose sur l'utilisation combinée d'un système expert centré-objet (SHIRKA) et des outils de l'interaction graphique.

L'aide apportée aux biologistes se concrétise à trois niveaux:

1. Le champ des tâches à accomplir par l'utilisateur doit être restreint aux préoccupations immédiates de celui-ci. Cet objectif passe par la réalisation d'une interface-utilisateur graphique. Celle-ci permet de percevoir le logiciel et d'interagir avec lui à un niveau conceptuel, à l'aide d'objets familiers.

2. Le large éventail des méthodes numériques ou formelles ainsi qu'une vaste "base de modèles" mis à la disposition du biologiste doivent être accompagnés des connaissances relatives à leur emploi (connaissances d'ordre méthodologique, connaissances concernant la sélection et la construction d'un modèle, connaissances sur le choix d'une méthode). L'intégration de ces objets et de l'expertise au sein d'une même base de connaissances permet un mode de fonctionnement assurant:

   - une automatisation des choix purement techniques,
   - la suggestion de différents choix possibles à chaque étape de la modélisation,
   - l'explication du raisonnement conduisant à la suggestion de ces choix,
   - l'exposé des répercussions des différents choix possibles.

3. La prise de décision et l'interprétation des résultats sont d'autant plus efficaces qu'il y a peu d'intermédiaires entre les modèles mentaux de l'utilisateur et l'information présentée à l'écran. La représentation des données résultant des méthodes de calculs doit donc privilégier la perception visuelle.
   On peut envisager la généralisation de cette approche à toute discipline mettant en œuvre un ensemble de méthodes de calcul scientifique sur des objets à fort contenu sémantique; l'économétrie par exemple.

1. Introduction

Dans le processus de construction d'un modèle mathématique, trois étapes jouent un rôle central. Ces étapes sont globalement suivies de façon séquentielle mais les nombreux essais/erreurs et raffinements successifs du modèle imposent de fréquents retours vers les étapes antérieures.

1.1. L'analyse du système et la formulation mathématique

Cette phase du processus recouvre le passage des hypothèses sur la structure et le fonctionnement du système à un ensemble d'équations répondant aux objectifs. Cette étape repose sur l'expérience et une bonne connaissance de la bibliographie. Elle peut éventuellement faire appel aux méthodes de l'analyse des données.

1.2. L'analyse du modèle

Dans le cadre des modèles à base d'équations différentielles, cette étape implique l'utilisation de méthodes numériques sophistiquées visant à:

* la simulation (intégration et tracé des solutions).
* la recherche des états stationnaires et l'étude de leur stabilité.
* la recherche des solutions périodiques.
* la recherche des points de bifurcation.
* l'analyse de la sensibilité du modèle.

Chacun de ces points nécessite des compétences mathématiques dans le domaine des équations différentielles.

1.3. L'identification des paramètres du modèle

Cette opération nécessite
* l'étude de l'identifiabilité du modèle .
* le choix d'un critère et le calcul d'un optimum.
* l' interprétation des résultats.

qui relèvent de l'analyse numérique, de l'automatique et de la statistique.

Chacune de ces étapes met en œuvre des compétences spécifiques qui ne sont pas du ressort du biologiste . A un degré moindre, un chercheur ayant les compétences requises est souvent géné, voire rebuté, par les problèmes informatiques posés par la mise en œuvre ou l'utilisation informatique des techniques de calcul.

Après un état des lieux de l'utilisation de l'informatique dans le cadre de la modélisation en biologie, nous présentons ici la philosophie directrice, les solutions techniques retenues et les principaux axes de recherches pour la constitution du logiciel d'aide à la modélisation EDORA. L'objectif principal de ce logiciel est d'impliquer d'avantage l'utilisateur dans la construction de ses modèles en le libérant des tâches trop techniques ou secondaires.

2. L'état des lieux

Parmi les outils cités au paragraphe précédent, les méthodes de calcul numérique ont une importance essentielle, avec comme corrolaire, l'utilisation intensive de programmes informatiques dans une grande partie du processus de modélisation.

L'utilisation de ces programmes s'insère au sein d'un cycle Reflexion/Action général à toute discipline mettant en œuvre des méthodes de calcul scientifique.

Ce cycle peut se résumer ainsi:

1. Analyser le problème en termes de sous-objectifs impliquant des méthodes de calcul.

2. Pour chaque sous-objectif faire:

    2.1. Choisir une méthode et fixer ses options en fonction
        - du sous-objectif
        - de la nature des données fournies par les sous-objectifs antérieurs
    2.2. Eventuellement mettre les données sous la forme requise par la méthode
    2.3. Faire les calculs
    2.4. Interpréter les résultats fournis par la méthode
    2.5. Répondre aux questions:
        - le sous-objectif est-il atteint ?
        - quel est le prochain sous-objectif ?

Actuellement, l'aide fournie par l'outil informatique se concentre sur le plan calculatoire (2.3), et déborde sur l'interprétation(2.4) grâce à l'introduction des aides graphiques à l'interprétation développées surtout en analyse des données (**Auda**, 1983). Les aspects pratiques de la mise en forme des données (2.2) peuvent également être facilités par l'emploi de l'informatique (gestion de fichiers par exemple).

On peut alors se poser deux questions:

1. L'aide actuellement apportée par l'informatique est-elle satisfaisante pour le biologiste? Par satisfaisant on entend le fait que l'aide est suffisante pour limiter l'intervention des mathématiciens et informaticiens aux seuls points délicats.

2. Peut-on envisager d'étendre cette aide aux autres points du cycle Action/Reflexion, habituellement non pris en compte par les logiciels ?

Les logiciels susceptibles d'intervenir en modélisation se caractérisent par une grande variété dans leur forme et leur fonction:

2.1. Les logiciels "sur mesure"

La plupart des partenaires consultants du biologiste réalisent des programmes spécifiques au fur et à mesure des besoins. Eventuellement, une concertation au sein d'un même laboratoire permet de définir des normes de programmation qui autorisent la réunion de ces programmes sous la forme d'une programmathèque ouverte à chacun.

Malgré son efficacité locale et sa grande souplesse pour des utilisateurs "méthodologistes", cette démarche possède un certain nombre d'inconvénients:

Ces programmes sont isolés les uns des autres: la poursuite d'un objectif donné nécessite l'emploi successif d'un certain nombre d'entre eux. L'utilisateur doit donc posséder la connaissance du fonctionnement de chacun d'eux et, bien souvent, de la forme informatique des données qui leur sont soumises.

Ces préoccupations informatiques, sans rapport avec l'objectif fondamental de l'utilisateur, perturbent le rythme de sa réflexion et handicapent la progression de sa recherche.

Enfin, la rusticité de l'interface-utilisateur de ces programmes limite leur emploi dans un cercle restreint autour de leurs concepteurs. Ils ne sont donc pas diffusables et ne peuvent apporter une aide au biologiste qu'au travers de l'intervention d'un consultant.

2.2. Les logiciels spécialisés

On trouve un certain nombre de logiciels spécialisés dans une tâche très précise.

Par exemple:

Intégration numérique (LSODA : **Hindmarsh**, 1983)
Identification (bibliothèques HARWELL et NAG)
Calcul formel (MACSYMA, REDUCE)
Analyse de sensibilité (SAP: **Arnaud**, 1984, HEQS: **Derman**, 1985)
Analyse des bifurcations (AUTO: **Doedel**, 1981)
Gestion de bases de données
Représentations graphiques (GRAPHIQUE: **Auda**, 1984)

Ces programmes sont, dans leur spécialité, très généraux et nécessitent donc pour leur mise en œuvre, le réglage d'un grand nombre de paramètres et le choix d'une ou plusieurs options. Ces choix sont, la plupart du temps, motivés par des arguments techniques, qui ne relèvent pas du domaine de compétence de l'utilisateur biologiste.

En outre, tout utilisateur doit naviguer entre les mêmes écueils que ceux des programmes "sur mesure" : apprentissage du mode d'emploi et de ses finesses, absence de communication entre les logiciels ,d'autant plus graves que ceux-ci tournent souvent sur des machines différentes.

2.3. Les logiciels intégrés

Ces derniers sont conçus dans le but d'intégrer dans un même environnement informatique un grand nombre d'outils nécessaires pour mener à bien la conception d'un modèle (DYNAMO, COSMOS, CSSL, CSMP, MLAB, BIOMOD, MODULECO...).

Ces logiciels s'appuient généralement sur un langage qui permet à l'utilisateur de décrire son modèle en termes mathématiques, et sur un langage de commandes qui permet d'effectuer des opérations ou d'appliquer des méthodes sur les données et les modèles.

Certains de ces logiciels intègrent un système de gestion de bases de données et un module pour la représentation graphique des données.

Malgré l'apport positif de ce type de logiciels qui déchargent l'utilisateur de la plupart des problèmes informatiques, on peut leur adresser plusieurs critiques:

1) Leurs concepteurs ont négligé les méthodes employées dans un certain nombre d'étapes de la modélisation: analyse du modèle (stabilité, sensibilité ...), calcul symbolique, analyse dimensionnelle, ... ;

2) Cette situation est agravée par le fait que ces logiciels sont figés: il est difficile de les modifier ou d'étendre leurs fonctionnalités. De plus, Ils sont souvent spécialisés dans un domaine d'application (par exemple l'économétrie : MODULECO) ou orientés vers une conception particulière de la modélisation (par exemple la dynamique des systèmes: DYNAMO);

3) Aucune information n'est fournie quant à l'utilisation des méthodes disponibles (pour l'intégration, l'identification,...) suivant le contexte (type de modèle, propriété des données...);

4) Malgré un effort récent, peu de logiciels intégrés sont réellement interactifs et ne possèdent pas d'interface-utilisateur autorisant une communication homme/machine efficace. Les langages de commande et les langages de description de modèle à la syntaxe plus ou moins complexe les rendent d'un abord difficile.

2.4. Conclusion

Cette revue rapide met en lumière trois points principaux:

1) Dans la majorité des cas, l'aide apportée par l'informatique est insuffisante pour une décentralisation des tâches de la modélisation vers les expérimentateurs.

Les logiciels les plus répandus exigent de leurs utilisateurs des connaissances mathématiques et informatiques qui sortent de leur domaine de compétence, et nécessitent un investissement important pour l'apprentissage de leur fonctionnement.

Cet état de fait, joint au manque de diffusion des logiciels, peut entrainer:

- Une sous exploitation des données biologiques. La difficulté de l'accès aux logiciels n'encourage pas les biologistes à aller au delà de quelques analyses de routine standards. Des traitements aussi simple que la représentation graphique des données de base ne sont souvent pas réalisés.

- Un nombre important d'utilisations abusives ou de sous-emplois des méthodes disponibles. L'utilisateur n'est pas guidé pour le choix d'une méthode, l'interprétation des résultats et la succession des étapes à accomplir. Des résultats incorrects peuvent ainsi être obtenus sans que l'utilisateur ait les moyens de s'en rendre compte.

La solution de ce problème passe nécessairement par l'intégration au logiciel des connaissances relatives à l'utilisation des méthodes et de la sémantique attachée aux objets qu'elles manipulent (**Rechenmann** 1985, **Saurel** 1985, **IFIP** 1985).

D'autre part, il est essentiel de bâtir des interfaces-utilisateurs permettant aux biologistes d'interagir avec le logiciel de la façon la plus simple et la plus naturelle possible.

2) L'impossibilité de faire évoluer ou de modifier les logiciels pourrait être compensée par l'introduction de systèmes de gestion de programmes permettant d'organiser et de structurer une base de programmes de calculs.

Seuls quelques logiciels sont conçus dans cet esprit (CS: **Dawson et al.** 1980, S: **Becker et al.** 1984), mais ils n'autorisent pas l'intégration et le traitement des connaissances sur les données et les méthodes.

3) Certains aspects de la modélisation sont souvent ignorés du fait de l'absence de logiciels permettant de les traiter. On peut citer, par exemple:
- l'analyse du modèle;
- la gestion des versions successives d'un modèle lors de sa conception;
- l'aide à la formulation mathématique à partir d'hypothèses biologiques.

3. Les Outils

Un certain nombre de concepts nouveaux sont en passe de provoquer de profonds bouleversements dans la manière de percevoir et d'utiliser l'informatique. Parmi ceux-ci, on peut citer les concepts de système à base de connaissances et de poste de travail.

3.1. Les systèmes à base de connaissances

Pour décrire de grandes quantités de connaissances et pour les faire évoluer facilement et rapidement, il est nécessaire de disposer d'un formalisme de représentation. Le formalisme informatique classique, sous la forme de programmes écrits dans un quelconque langage aussi évolué soit-il, ne résoud pas la question. Les programmes obtenus sont gros et complexes, difficiles à corriger et à maintenir.

Pour ces raisons, les chercheurs en Intelligence Artificielle développent depuis de nombreuses années des modèles de représentation bien adaptés au développement de systèmes utilisant de grandes quantités de connaissance. Ces systèmes sont dits experts quand leur ambition est de reproduire le comportement d'un expert pour une catégorie de tâches donnée: diagnostic médical ou financier, surveillance d'unités de production, etc.

• *3.1.1. Les règles de production*

Le modèle de représentation actuellement le plus apprécié utilise les règles de production. Une règle de production est de la forme:

SI condition ALORS action

La partie condition est composée d'une ou de plusieurs prémisses portant sur des faits élémentaires dont la véracité est soit connue a priori en tant qu'hypothèse, soit inférée à l'aide d'autres règles de la base de connaissances. L'action la plus fréquemment rencontrée en partie droite consiste à ajouter de nouveaux faits lorsque la règle est déclenchée, c'est à dire quand les prémisses en partie gauche sont toutes vérifiées.

Il existe plusieurs modes d'exploitation de ces règles suivant que l'on parte de faits connus afin d'en inférer de nouveaux ou qu'au contraire on se fixe des buts à vérifier en recherchant les faits correspondants. Dans tous les cas, chaînage avant, chaînage arrière ou mixte, le travail est effectué par le moteur d'inférence dont l'algorithme consiste principalement à boucler en étudiant l'état de la base des faits et en déclenchant les régles sélectionnées.

Les avantages de cette représentation des connaissances sont désormais bien connus et tournent autour de la notion de modularité. Une règle n'incorpore qu'une petite quantité de connaissance. Elle est en théorie indépendante des autres. Les règles ne communiquent qu'à travers la base de faits. Le développement et les modifications de la base en sont grandement facilités.

3.1.2. Les règles de production et l'aide à la modélisation

L'utilisation d'une représentation des connaissances à l'aide de règles de production a été illustrée par de Swaan Arons (1983). L'exemple simple considéré consiste à aider à définir le modèle mathématique d'un système physique élémentaire, constitué d'une masse accrochée à un ressort, compte tenu des hypothéses a priori sur le système, telles que l'existence d'une force de friction ou l'importance de la masse du ressort.

Exemple

Règle 13:

SI

le ressort ne satisfait pas la loi de Hooke
le ressort est de masse nulle
il n'y a pas de force de friction
il n'a pas de force extérieure

ALORS

le modèle est $\frac{d^2x}{dt^2} + \frac{f(x)}{M} = 0$

Sur les vingt règles que comporte la base, douze possèdent en conclusion, comme la règle de l'exemple, la forme du modèle retenu. Ces règles définissent en fait une classification des modèles possibles d'un système masse-ressort. Cette classification est hiérarchique et un modèle d'un niveau traduit des hypothèses communes aux modèles plus spécifiques. L'utilisation de règles fait disparaître, par dispersion, cette idée de classification.

En fait, cette représentation des connaissances répond très mal aux besoins d'un système d'aide à la modélisation; elle ne permet pas de représenter explicitement les objets manipulés: modèles, équations, coefficients, données. Cette représentation explicite est indispensable. En effet, moteur d'inférence et programmes de manipulation doivent communiquer. Le moteur utilise des faits élémentaires, linéarité d'une équation ou signe de la partie réelle d'une valeur propre, qui sont en fait des propriétés extraites des objets par des programmes de calcul symbolique ou numérique. Inversement, le choix d'une méthode algorithmique adéquate est sous la responsabilité du moteur d'inférence. Il doit donc exister également une description explicite des méthodes disponibles: leurs conditions d'application et le code informatique correspondant. La solution qui consisterait à définir trois bases séparées, connaissances, objets et méthodes, pose d'importants problèmes quant au maintien de leur cohérence mutuelle. Pour ces raisons, une représentation de connaissances dite "centrée-objet" a été retenue.

3.1.3. Les représentations centrées-objet: les schémas

Les représentations de connaissance centrées-objet dérivent essentiellement des travaux de Minsky (1975) sur les "frames" et des travaux sur les réseaux sémantiques (81). En fait, l'article de Minsky ne constitue qu'un cadre de recherche et n'offre aucune proposition technique de réalisation des idées exposées. Il faut y voir la raison des nombreuses interprétations actuelles de la notion de "frame" qui n'ont retenu de ce cadre que les aspects les plus immédiats.

Pour satisfaire les exigences du projet EDORA, une représentation centrée-objet particulière a été développée. L'entité de description y est appelée **schéma**. Un schéma décrit aussi bien une classe d'objets qu'un objet particulier, représentant d'une classe. Le terme d'objet doit être pris dans son acception la plus large: ce peut être un objet physique ou conceptuel, une situation ou un contexte, une méthode algorithmique.

Un **schéma** est défini par un nom et la liste de ses attributs. Un **attribut** est défini par un nom et une liste de facettes. Il peut représenter une propriété ou un lien avec une autre classe. Les **facettes** permettent de définir le type de l'attribut, c'est à dire le domaine général de ses valeurs. Ce type peut être précisé par la donnée de prédicats devant être vérifiés pour toute valeur de l'attribut auquel ils sont attachés. D'autres facettes décrivent les moyens d'obtenir la valeur inconnue d'un attribut d'un représentant de la classe décrite par le schéma. Il est ainsi possible de définir une valeur fixe, de déterminer la valeur en l'extrayant d'un objet dont une définition partielle est spécifiée, de calculer la valeur en faisant appel à une procédure ou encore de retenir une valeur par défaut quand les moyens précédents ont échoué.

Ainsi, quand un schéma décrit une méthode algorithmique, ses attributs définissent les paramètres d'entrée et de sortie: leurs types, les conditions que leurs valeurs doivent vérifier pour que l'appel puisse avoir lieu, ainsi que les moyens de déterminer leurs valeurs si la forme d'appel ne les fournit pas.

Les schémas permettent donc la description des classes des objets manipulés dans EDORA ainsi que la description des méthodes algorithmiques au sein d'une seule base. En effet, un modèle particulier y sera un représentant de la classe modèle. Ses attributs seront par exemple la forme mathématique, différentielle et intégrée, la position de ses asymptotes et la valeur de ses paramètres, autant d'informations exploitables et dérivables par des méthodes procédurales. Un autre attribut permettra de faire le lien avec des modèles voisins. Toute la connaissance attachée à ce modèle pourra être trouvée dans le schéma de sa classe.

La conception de cette base est grandement facilitée par sa structure de treillis, sur laquelle un mécanisme d'**héritage** permet à un schéma de classe donné d'hériter des éléments de description des schémas plus généraux qui le dominent. Par l'héritage, les modifications effectuées sur une classe donnée sont de fait propagées à toutes ses sous-classes. De plus, à partir d'une certaine taille, l'adjonction dans la base d'un nouveau schéma de classe se ramène à la définition d'une variante d'un ou de plusieurs schémas existants. Enfin, ce treillis peut être employé à des fins de classification. Un objet, créé dans une classe donnée, peut être placé convenablement dans les classes inférieures s'il satisfait les conditions qui y sont attachées. La figure 1 fournit un exemple de treillis de classes tel qu'il pourrait être défini pour un petit sous-ensemble de l'algébre linéaire.

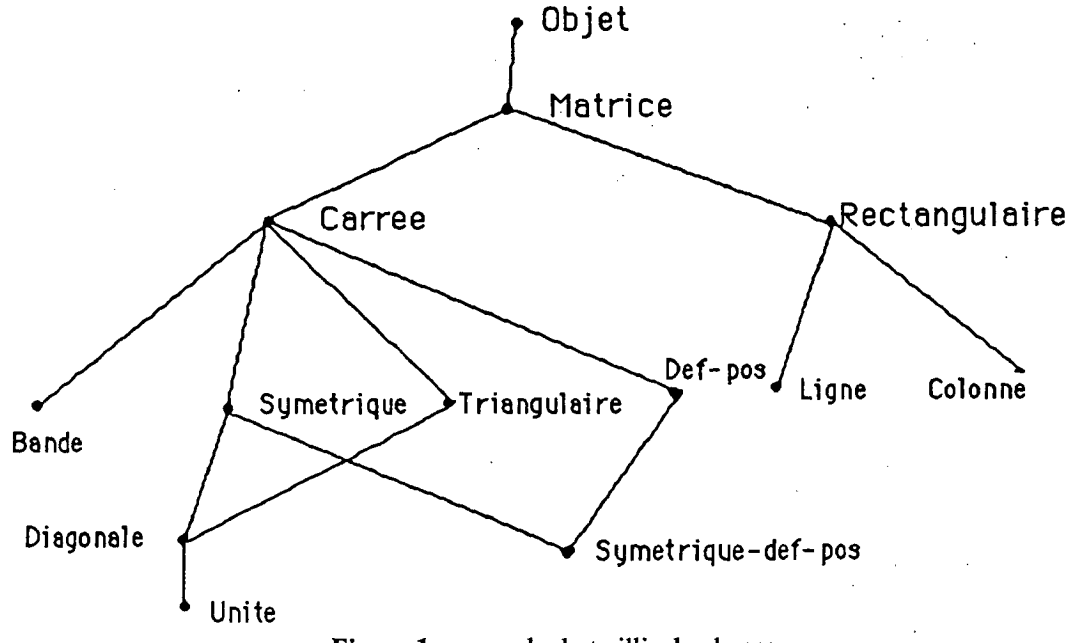


Figure 1 - exemple de treillis de classes.

3.2. Les postes de travail

La notion de poste de travail (**MacDonald et Pedersen** 85) mêle des aspects matériels et logiciels. Un poste de travail informatique (Appolo, Sun, Symbolics ...) est un ordinateur dédié à un seul utilisateur, possédant une grande puissance de calcul et offrant un très grand espace mémoire. La plupart de ces matériels possèdent en outre un écran bitmap haute résolution, des dispositifs de pointage (souris, tablette à digitaliser ...) et le logiciel associé (gestion des évènements, graphique, multifenêtrage ...).

Cette combinaison permet un langage graphique et naturel pour communiquer entre utilisateur et ordinateur, qui à son tour, conduit à la conception d'environnements de calcul plus puissants, sophistiqués et faciles à utiliser.

3.3. Le projet EDORA

Le projet EDORA s'appuie sur le constat de l'insuffisance relative des logiciels existants et sur l'apparition de voies nouvelles ouvertes par les progrès récents de l'informatique. Son objectif est de fournir aux biologistes un logiciel incluant la majorité des outils nécessaires à la modélisation et capable de les guider à toutes les étapes de la conception d'un modèle. Le noyau central de ce logiciel est constitué d'un système expert travaillant sur une base de connaissances centrée-objet.

Il convient de faire plusieurs remarques préliminaires afin de tempérer les réactions dubitatives ou exagérément optimistes suscitées par un projet aussi ambitieux.

• Les termes d'intelligence artificielle et de systèmes experts sont très en vogue, ce qui est la source de bien des abus ou d'espoirs non fondés. Ce projet ne vise pas à automatiser le processus de modélisation mais, au contraire, à libérer l'utilisateur des tâches secondaires pour mettre en lumière les choix importants qui ne peuvent être pris que par lui. D'autre part, l'intérêt du système expert, plus encore que ses fonctions "intelligentes" qui peuvent toujours être réalisées au moyen de programmes classiques, réside dans son rôle structurant permettant d'organiser et de maintenir le logiciel avec l'intervention minimum des informaticiens.

• Malgré l'aide que le biologiste peut attendre d'un tel logiciel, certaines compétences requises pour la modélisation (notamment l'analyse fine du comportement d'un modèle) sont si spécialisées qu'elles nécessitent l'intervention du mathématicien ou du statisticien. Le projet EDORA a pour but de limiter cette intervention et, lorsque celle-ci est requise, de mettre à la disposition des partenaires des outils de travail communs efficaces.

• Au travers de l'objectif à plus ou moins long terme d'un produit fini, le projet EDORA permet de catalyser les efforts au niveau de la recherche méthodologique, de la remise en cause des pratiques et de la formulation des connaissances utilisée par les modélisateurs.

Une réflexion sur les différents types d'aide à apporter aux biologistes a permis de dégager trois axes de recherche principaux :

- l'aide à la prise de décision,
- l'aide par une meilleure communication homme/machine,
- l'aide à la formulation mathématique du modèle,

4. Trois axes de recherche

4.1. L'aide à la prise de décision

Le cycle action/réflexion suivi par le modélisateur comporte de nombreux choix et prises de décision: choix d'une méthode, choix d'un modèle, choix de l'étape suivante...

Souvent désarmé devant la multiplicité de ces choix, l'utilisateur doit être guidé et conseillé à chaque étape: le logiciel doit donc pouvoir prendre des décisions dépendant du contexte et de la nature des modèles et des données manipulés.

Dans certains cas, on peut envisager d'automatiser la prise de décision; par exemple un choix purement technique comme celui d'une méthode d'intégration peut être effectué en fonction de la nature des équations.
Cependant, cela n'est ni souhaitable ni possible dans la plupart des cas et le logiciel devra alors être capable de présenter de la façon la mieux adaptée l'information permettant à l'utilisateur de faire ce choix, ou de lui proposer un stratégie d'essais/erreurs permettant de s'approcher progressivement d'une solution. Par contre, l'utilisateur doit pouvoir demander une explication du raisonnement aboutissant à ces suggestions.

Un tel mode de fonctionnement, illustré par la maquette en cours de réalisation et décrite plus loin, implique que le logiciel ait accès à une certaine forme d'expertise couvrant: la connaissance des propriétés des objets traités par les méthodes, la connaissance des conditions d'application des méthodes, et la connaissance des buts visés par l'utilisateur.

Cet objectif fait clairement appel aux techniques de l'intelligence artificielle.

Rappelons donc ici que le logiciel développé dans le cadre du projet EDORA sera construit autour d'un système expert travaillant à partir d'une représentation des connaissances centrée-objet (SHIRKA: **Rechenmann** 1985). Cette forme de représentation des connaissances permet la description, l'organisation et la gestion des objets, des méthodes et de l'expertise au sein d'une même base de connaissance.

SHIRKA assure à la fois la détermination des propriétés des objets, le choix des méthodes appropriées, l'exécution des modules qui les réalisent, l'évaluation des résultats et enfin la mémorisation des propriétés des objets résultants afin d'éviter, dans le cas d'enchainements de méthodes, des inférences ultérieures inutiles. Les objets doivent donc apparaitre dans la base de connaissance, non seulement à travers leurs propriétés, mais aussi à travers leur représentation informatique opérationnelle (par exemple: pour une matrice, la valeur de ses éléments). De même, à toute description d'une méthode doit être associé le module informatique qui l'exécute.

La première application envisagée concerne le choix et l'identification d'un modèle de croissance à partir de données sur la croissance d'un organisme ou d'une population. Conformément aux idées présentées précédemment, il s'agit plus de guider ce choix que de l'automatiser. En fait, plusieurs modèles peuvent être éventuellement retenus pour un même jeu de données.

Plusieurs étapes sont ainsi prévues:

- Sélection d'un premier modèle adéquat à partir de l'allure générale du nuage des points expérimentaux.
- Estimation initiale des paramètres par une méthode dépendante du modèle retenu.
- Identification proprement dite du modèle.

- Sélection par l'utilisateur d'un autre modèle parmi ceux conseillés. Ultérieurement, le choix d'un autre modèle pourra être guidé par les résultats de l'identification à travers l'exploitation de statistiques non paramétriques.
- Sélection a posteriori du meilleur modèle parmi ceux essayés.

Dans ces conditions, la base doit contenir des connaissances sur les différents modèles possibles. Actuellement, les classes de modèles retenus sont les suivantes: exponentiel, Gompertz, Logistique, Logistique généralisée, monomoléculaire, Monod et Kostitzin, certaines classes possédant des variantes. Chaque classe est décrite par un ensemble de propriétés et d'attributs: formes différentielle et intégrée du modèle mathématique quand elle est connue, existence d'asymptotes inférieures et supérieures, paramètres, et surtout relations avec les classes considérées comme voisines du point de vue de leur adéquation aux données considérées.

L'adéquation entre un jeu de données et les classes de modèles se fonde, conformément à la pratique en ce domaine, sur l'allure générale des observations. En effet, les données étant visualisées sous la forme de graphiques adéquats, le biologiste fait rapidement le rapprochement avec les modèles de croissance qu'il connait et il n'expérimente l'estimation que sur les modèles ainsi sélectionnés.

Cependant, ce mode de travail n'est pas directement accessible à un programme informatique auquel il manque fondamentalement cette perception globale des formes. Là où le biologiste voit immédiatement une suite de points en forme de cloche, il faudrait imaginer un système de reconnaissance de forme dont la complexité dépasserait largement celle du système envisagé ici. Il est donc naturel d'avoir recours aux compétences de chacun: la machine affiche le graphique représentant les données étudiées et le biologiste décrit à l'aide de mots-clés la forme qu'il perçoit. A l'aide de cette description, et en utilisant sa connaissance sur les modèles de croissance, le système peut déterminer une première liste des modèles à envisager.

4.2. L'aide par une meilleure communication homme/machine

L'ordinateur a pour fonction de libérer l'utilisateur des tâches mécaniques. Ce but semble atteint en ce qui concerne le calcul numérique et le stockage massif des données. Cependant, trop de tâches fastidieuses viennent encore briser le fil de la reflexion de celui-ci : langages, options et manipulations de fichiers par exemple.

En outre, comme le soulignent (**McDonald et Pedersen** 1985), bien que les contraintes informatiques aient été largement réduites, la nature "rigide" de la plupart des logiciels scientifiques fait que leurs utilisateurs ont perdu la flexibilité du mode de travail papier/crayon.

Les concepts de l'interaction graphique (**Foley et al.** 1984, **Averill** 1984) tels qu'ils sont mis en œuvre sur certains postes de travail (icones, menu déroulant, multifenètrage...) peuvent constituer le point de départ d'une recherche visant à la conception d'interfaces-utilisateurs limitant ces problèmes, dans la mesure où elles sont accompagnées d'une réflexion méthodologique.

Une telle recherche s'appuie sur le fait qu'il existe un lien très fort entre la forme externe que revêt un programme (ses entrées/sorties) et la façon dont l'utilisateur va intégrer ce programme dans son schéma Action/Reflexion.

Pour que cette intégration soit la plus efficace possible, il est nécessaire de limiter le nombre d'intermédiaires entre les modèles mentaux de l'utilisateur et les entrées/sorties du programme (**Kay**, 1984).

Une interface-utilisateur doit donc être construite à partir d'un modèle conceptuel, environnement imaginaire dans lequel l'utilisateur est capable d'agir à l'aide d'objets familiers. Ainsi, l'environnement de bureau du MacIntosh est un modèle conceptuel de son système d'exploitation.

Dans cet esprit, on peut envisager de transformer la pratique d'un certain nombre de techniques utilisées en calcul scientifique et en modélisation en particulier. De nombreuses voies de recherche sont ouvertes dans ce domaine.

Une voie prometteuse concerne la réalisation de véritables cahiers de brouillon informatiques (CABRI: **Benzaken** 1986) ayant pour but de rendre facile et naturel l'observation des objets et les traitements de base effectués sur ceux-ci (graphiques, statistiques, intégration numérique, calcul formel ...). Les algorithmes utilisés doivent se fondre dans une interface-utilisateur permettant une analyse exploratoire de type papier/crayon (**Tukey** 1979) mais bénéficiant de la puissance de l'ordinateur. Ce dernier tend ainsi à devenir une extension de la volonté de l'utilisateur qui n'hésite alors plus à se poser des questions du type "Que se passe-t-il si ... ?" .

De façon générale, la représentation conceptuelle du fonctionnement d'un algorithme et le choix des représentations graphiques de ses résultats ,permet de s'assurer de la justesse de ce que perçoit l'utilisateur, ce qui constitue une forme d'intégration de l'expertise dans le logiciel.

De telles interfaces ne relèvent pas du gadget ni du luxe; elles doivent permettre un travail plus efficace, plus simple mais ne se limitent pas à ces aspects. De par la nature et la qualité du dialogue qui s'établit avec l'utilisateur, elles doivent entrainer l'évolution des pratiques et une amélioration qualitative dans la compréhension des phénomènes étudiés.

A l'actif du projet EDORA, on peut citer deux réalisations, pour l'instant déconnectées de toute base de connaissance, mais qui préfigurent les interfaces utilisateur du logiciel à venir.

4.2.1. Le programme DYNAMAC

Ce programme permet l'étude graphique interactive des systèmes différentiels ou récurrents à deux dimensions, sur MacIntosh. L'utilisateur commence par décrire ses équations à l'aide d'un mini-langage. L'étude du système repose ensuite sur le tracé graphique de ses solutions dans le plan de phase.

Le mécanisme de base consiste à choisir des conditions initiales en se positionnant dans le plan de phase à l'aide de la souris, puis à cliquer pour lancer l'intégration (figures 2 et 3). De nombreuses options complémentaires permettent le choix d'une méthode d'intégration, le tracé du champ de vecteurs, le tracé des isoclines, la représentation des courbes d'évolution en fonction du temps ainsi que la localisation et la détermination des points d'équilibre.

4.2.2. Le programme CROISSANCE

Ce dernier est conçu pour l'identification des modèles de croissance à une variable d'état. Outre les calculs classiques (estimation initiale, puis estimation par la méthode de Gauss-Marquardt, il permet une "identification à main levé" qui illustre bien la notion d'algorithme papier/crayon. Chacune des courbes de croissance est caractérisée par des points de contrôles dépendant des paramètres et dont l'emplacement dans le plan variable-temps définit entièrement la forme de la courbe. Par exemple, pour le modèle logistique, ces points de contrôle sont le point d'inflexion, l'ordonnée à l'origine et les assymptotes inférieures et supérieures (figure 4).

Pour ajuster la courbe sur les points expérimentaux, il suffit de déplacer l'un des points de controle à l'aide de la souris. La courbe se déforme alors, et la modification de forme est immédiatement répercutée sur les valeurs des paramètres.

4.3. L'aide à la formulation mathématique d'un modèle

Une facette importante de la modèlisation n'a encore donné lieu qu'à peu d'applications informatiques: il s'agit de l'aide à la formulation mathématique à partir d'un ensemble d'hypothèses (COSMOS: **Hamrouni** 1979 dans le cas des modèles à compartiments).

Cette aide peut être fournie à deux niveaux:

- la recherche dans une base de modèles, d'un ou plusieurs modèles répondant à certaines conditions et adaptées à un domaine particulier (avec références bibliographiques),

- la construction progressive du modèle à l'aide des hypothèses sur sa structure et son fonctionnement, du domaine d'application et éventuellement de la forme des données expérimentales.

Dans les deux cas, cette approche nécessite l'existence

* d'un ou plusieurs formalismes non mathématiques permettant à l'utilisateur de décrire le modèle en termes de structure et de fonctionnement,

* d'une structure de représentation informatique des modèles permettant l'enregistrement de leurs caractéristiques,

* d'algorithmes et d'heuristiques permettant de passer progressivement du formalisme de description au formalisme mathématique, par combinaison de sous-structures de base correspondant à des situations biologiques élémentaires (**Zeigler** 1979).

De nombreux formalismes plus ou moins généraux et plus ou moins adaptés à la biologie ont été proposés, dont la plupart s'appuient sur des représentations graphiques à base de graphes:

- Modèles à compartiment
- Graphes causaux (**Roberts** 1983)
- Flow Graphs (**Wiitanen** 1976)
- Bond Graphs (**Karnopp** 1975)
- Energy Circuit langage (**Odum** 1983)
- Equations de type chimique (**Pavé** 1980)

Le passage du formalisme de description au formalisme mathématique peut être effectué de façon progressive, les étapes étant guidées par le logiciel. On peut ainsi envisager une stratégie telle que:

- définir les variables
- spécifier l'existence d'interactions entre ces variables
- définir plus précisément la nature de ces interactions.

Cette approche soulève de gros problèmes qui ne peuvent être résolus que par l'utilisation coordonnée
- de techniques de représentation et d'exploitation de la connaissance
- de méthodes de calcul symbolique (assemblage de primitives, simplification, re-paramétrages, analyse dimensionnelle)
- d'outils de la graphique interactive permettant d'envisager la réalisation d'un éditeur de modèle.

Le système SHIRKA semble a priori adapté à une telle approche. La représentation des connaissances centrée-objet permet la description des liens de dépendance entre modèles ou primitives, leur organisation sous forme hiérarchique en tenant compte des différents niveaux de description (biologiques et mathématiques) ainsi que la représentation des propriétés des modèles sous forme d'attributs (**Pavé et Rechenmann**,1985).

5. Conclusion

Les problèmes rencontrés par les biologistes face à la modélisation mathématique se retrouvent à l'identique dans d'autres disciplines, et particulièrement en économétrie. L'état des lieux y distinguerait de la même manière les logiciels sur mesure, les logiciels spécialisés et les logiciels intégrés, dont TROLL et Moduleco sont les représentants les plus actuels. Cependant, l'examen de la pratique de la modélisation économétrique révèle certaines différences importantes par rapport à la pratique du biométricien.

En tout premier lieu, l'économètre dispose d'observations sur lesquelles il n'a pas ou peu de contrôle, au contraire du biologiste qui peut parfois planifier des expériences in vitro afin d'améliorer ses données. Le processus de conception d'un modèle s'en trouve donc modifié. L'utilisation d'un formalisme de description des objets trouve ici son intérêt. Les étapes du processus et leur séquencement peuvent être décrites explicitement et facilement modifiées pour s'accomoder de telles différences.

Ensuite, la taille des modèles économétriques est de loin supérieure à la taille des modèles biologiques, trés souvent de deux ordres de grandeur. Dans ces conditions, certaines méthodes graphiques de visualisation ou de calcul doivent être complètement revues. De plus, l'interactivité possible est réduite par les temps de calcul impliqués. Il faut cependant remarquer que la construction de ces grands modèles n'est menée que par des organismes susceptibles d'en assurer le coût. Des modèles de petite taille peuvent être construits par exemple dans des centres de recherche universitaires.

Enfin, le nombre de méthodes disponibles, en particulier d'estimation des coefficients, est sans doute plus important qu'en biométrie. Une approche à base de connaissances apparaît donc d'autant plus justifiée.

D'autres différences, outre celles liées à la terminologie employée, existent:

– le formalisme mathématique majoritairement retenu en économétrie est le système d'équations aux différences. En fait, en biométrie, le formalisme mathématique est tantôt continu (equations différentielles), tantôt discret (equations aux différences), suivant les auteurs et les domaines d'application.

– compte tenu de la taille des modèles, une technique d'analyse intéressante consiste à tenter de construire un modèle réduit, ou maquette, des modèles dont la complexité gène la compréhension de leurs comportements. Les maquettes sont alors du même ordre de complexité que les modèles biologiques. Les techniques de manipulation et d'interaction proposées y sont sans doute applicables.

Il est donc probable qu'en ce qui concerne l'économétrie, une analyse semblable à celle qui vient d'être faite pour la biométrie amènerait à des conclusions similaires quant au développement souhaitable de logiciels d'aide à la modélisation.

Bibliographie

Auda Y (1983). Rôle des méthodes graphiques en analyse des données : application au dépouillement des enquêtes écologiques. *Thèse 3° cycle, Université Claude Bernard*, LYON 1.

Averill (1984). User Interface Guidelines. *Inside MacIntosh*, Apple Computer, 1984.

Benzaken C. (1986) Un éditeur de graphes simple, d'aide à l'enseignement et à la recherche. *Rapport technique IMAG N° 1*, Laboratoire LSD, GRENOBLE.

de Swaan Arons H. (1983).Expert Systems in the Simulation Domain. *Mathematics and Computers in Simulation* , XXV, 10-16

Foley J.D., Wallace V.L., Chan P. (1984). The human factors of computer graphics interaction techniques. *IEEE Comp. Graph. & Appl.*, 13-48.

Hamrouni M.K. (1979). Etude et développement d'un système informatique d'aide à l'élaboration de modèles en biologie. *Thèse 3° cycle*, Paris 6, 1979.

Karnopp D., Rosenberg R. (1975). *Systems Dynamics: a Unified Approach.* Wiley, New York.

IFIP WG 2.5 (1985). Problem solving environments for scientific computing. IFIP WG 2.5 working conference 4, INRIA Editeur, Sophia-Antipolis.

Kay A. (1984). Les logiciels. *Pour La Science*, °85, 14-22.

McDonald J.A., Pedersen J. (1985). Computing environments for data analysis.
I : Introduction
II : Hardware
SIAM J. Sci. Comput., 6, 4, 1004-1021, october 1985.

Minsky M. (1975). A Framework for Representing Knowledge, in "*The Psychology of Computer Vision*", P.H. Winston Ed., McGrawHill.

Mylopoulos J. (1981). An Overview of Knowledge Representation. *Sigplan Notices*, 16, 1.

Odum H.T. (1983). *Systems Ecology.* Wiley, New York, 1983

Pavé A., Rechenmann F. (1986). Computer aided modelling in biology, an artificial intelligence approach. "Artificial Intelligence in Simulation: State of the Art", Vansteenkiste and al. Eds, SCS publications, 18, .

Pavé A. (1980). Contribution à la théorie et à la pratique des modèles mathématiques pour l'analyse dynamique des systèmes biologiques. *Thèse Doct. ès Sciences*, Université Claude Bernard, LYON 1, 1980

Rechenmann F. (1985). SHIRKA: mécanismes d'inférence sur une base de connaissances centrée-objet. Cinquième Congrès AFCET-ADI-INRIA *"Reconnaissance des Formes et Intelligence Artificielle"*, Grenoble.

Rechenmann F. (1985). Représentation des connaissances dans les logiciels de calcul scientifique. In "*Informatique et Calcul, Computers and Computing*" P. Chenin, C. DiCrescenzo, F. Robert, Masson et Wiley ed., 1986. Actes du Congrès International "Le Calcul Demain", Grenoble, 2-6 décembre 1985.

Roberts N., Andersen D., Deal R., Garet M.,Shaffer W. (1983). *Introduction to computer simulation*. Addison-Wesley Publ. Comp..

Saurel C. (1985). Systèmes experts d'assistance aux réalisateurs de logiciels scientifiques. *Les systèmes experts et leurs applications*, Agence de l' Informatique Editeur.

Tukey J.W. (1979). Exploratory Data Analysis. Addison-Wesley, Reading, MA.

Wiitanen W. (1976).Modeling biological systems by means of flowgraphs.*Simulation*, 27, 1, 185-192, 1976.

Zeigler B.P. (1979). Structuring principles for multifaceted system modeling. In "*Methodology in systems modeling and simulation*" B.P. ZEIGLER and al. (Eds), North-Holland.

Logiciels cités

Arnaud M., Lammarre H. (1984). Un logiciel pour l'analyse de sensibilité des systèmes dynamiques. Rapport année spéciale ENSIMAG.

Auda Y. (1985). Logiciel graphique pour l'analyse des données (FORTRAN 77). *Laboratoire de Biométrie*, LYON 1, 107p.

Becker R.A., Chambers J.M. (1984). Design of the S System for Data Analysis. CACM, 27, 5, 486-495.

Dawson R., Klensin J.C., Yntema D.B. (1980) The Consistent System. *The American Statistician*, 34, 3, 169-176.

Derman E., Sheppard E.G. (1985). HEQS: a Hierarchical Equation Solver". *AT&T Technical Journal*, 24, 9, 2061-2096.

Doedel E.J. (1981). AUTO: A program for the automatic bifurcation analysis of autonomous systems.*Cong. Num.* 30, 265-284.

Hindmarsh A.C. (1983). ODEPACK, A systematized collection of ODE solvers. In "*Scientific Computing*", STEPELMAN et al. (Eds.), North-Holland Publ., Amsterdam, 55-64.

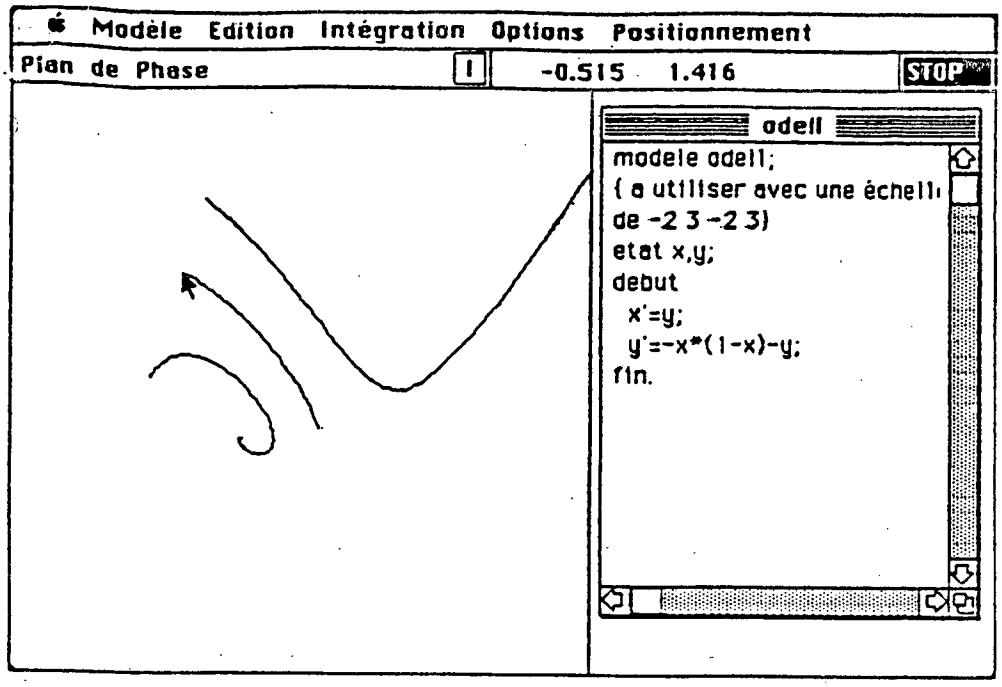


figure 1 - DYNAMAC : tracé de trajectoires dans le plan de phase.

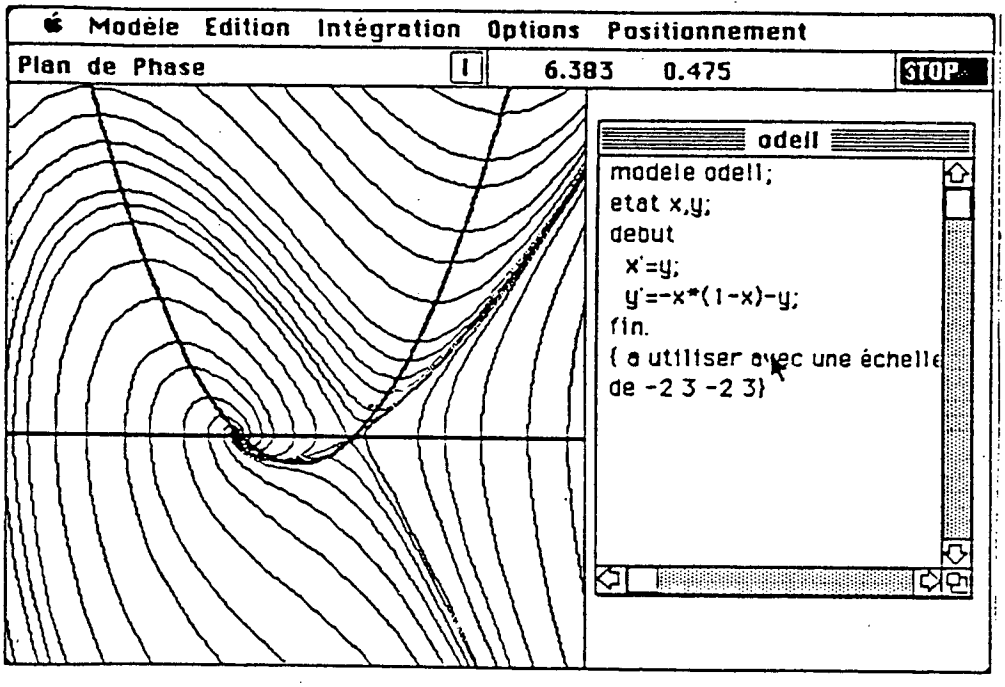


figure 2 - DYNAMAC : exemple de portrait de phase.

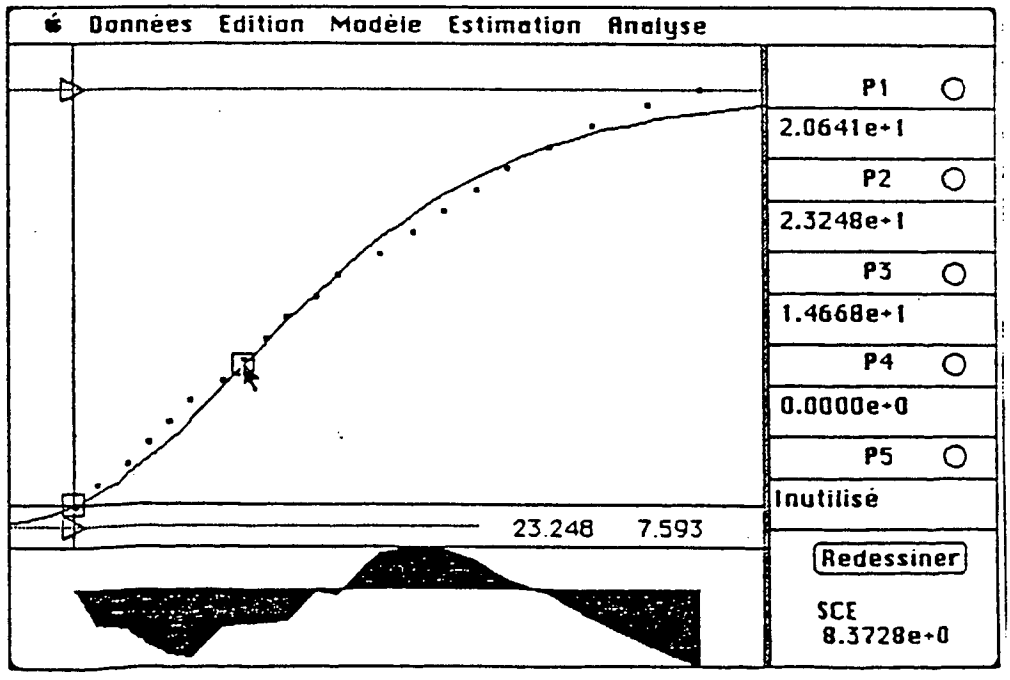


figure 3 - CROISSANCE : exemple du modèle de Gompertz. La forme de la courbe est pilotée par la position des quatres points de contrôle (carrés et triangles).

Vers l'intégration des objets symboliques et biologiques dans EDORA

EXPRESSIONS-MODELES- PROCESSUS-SYSTEMES

Christine PIERRET-GOLBREICH
INRIA - Rocquencourt
BP105 78153 Le Chesnay Cedex - France

Biologie

Rat musqué

Mathématiques

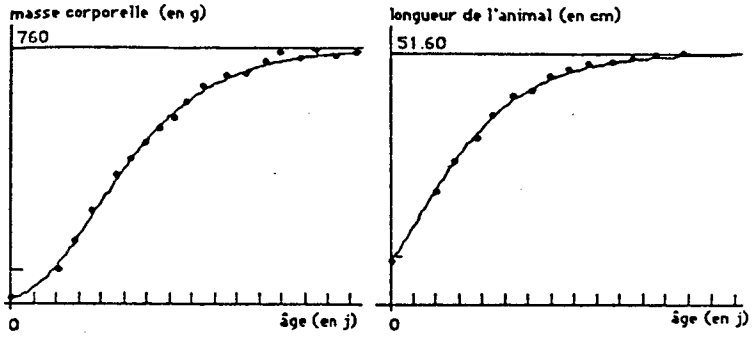


figure e3. masse corporelle et longueur de l'animal en fonction de l'âge pour le jeune rat

modèle de Gompertz

$$\frac{dx}{dt} = r\,x\left(1 - \frac{\ln x}{\ln K}\right)$$

modèle logistique

$$\frac{dx}{dt} = r\,x\left(1 - \frac{x}{K}\right)$$

CROISSANCE DU RAT MUSQUE

```
(def-sh
        '(exemple-gompertz
                (est-un (= modele-gompertz ))
                (variables (= ($taille-rat-musqué )))
                (paramètres (= (1.1 53)))
                (équations (= equ-dif1))
                (cond-ini (= 17))))
```

OBJETS SHIRKA

Les progrès récents de l'Informatique, notamment dans le domaine de l'Intelligence Artificielle, conduisent à concevoir des logiciels d'aide à la modélisation de type nouveau. L'objectif du projet Edora est de fournir aux biologistes un environnement informatique capable de les aider au cours de toutes les étapes d'un processus de modélisation: aussi bien lors de la formulation du modèle, que lors de sa simulation ou de sa validation. Un tel système doit donc à la fois incorporer les outils permettant de décharger l'utilisateur de tout effort de programmation ou de manipulation informatique de base et inclure les connaissances des domaines méthodologiques impliqués (calcul numérique, formel, statistiques...) ainsi que celles du domaine d'application. Un tel logiciel doit présenter les fonctionnalités des logiciels classiques de simulation tout en s'en distinguant par l'introduction d'importantes spécificités: par exemple, privilégier l'assistance au choix ou à l'utilisation de méthodes plutôt que multiplier les méthodes disponibles, prendre en compte les particularités des domaines d'applications plutôt qu'être conçu por une vaste gamme d'applications.

Les différentes expériences et réflexions menées dans le cadre du projet EDORA, ont permis de dégager les caractéristiques essentielles souhaitées pour le système: il doit permettre la coexistence et la gestion d'une base d'objets de natures diverses, d'une base de méthodes algorithmiques et d'une base de connaissance. Les utilisateurs potentiels du système ne sont guère supposés être particulièrement familiarisés à la modélisation, ou à l'informatique mais sont par contre beaucoup plus aptes à exprimer leur vue du problème en termes biologiques. Une attention particulière doit donc être portée au développement d'un formalisme permettant de capter les informations significatives formulées par le biologiste sur le système à modéliser. Ces priorités ont dicté un certain nombre de choix pour Edora notamment:

le mode de représentation de la connaissance retenu est basé sur la représentation centrée-objet.

les matériels visés sont de type Sun, Apollo, ou Macintosh illustrant le concept de poste de travail.

Ces orientations nécessitent d'intégrer au système un certain nombre d'outils spécifiques. Cet article porte sur les outils de manipulation symbolique : "calcul formel" minimal, manipulation de modèles. Il s'agit de montrer comment une technique privilégiée de l'Intelligence Artificielle, représentation de la connaissance centrée objet (Shirka), permet de définir les notions mathématiques usuelles : expression, fonction, équation-différentielle, modèle, tout en intégrant les connaissances biologiques de la situation étudiée.

I - INTRODUCTION

Le système EDORA utilise un mode de représentation de la connaissance centrée objet basé sur le concept de "frames" [11]. L'unité syntaxique de représentation est un "schéma" [16].

Un modèle pour EDORA a une signification précise: celle de modèle mathématique (au sens de la description mathématique d'un système dynamique pour l'automatique). On ne distingue pas différentes notions de modèle comme celles de "modèle formel" ou de "modèle physique" [5].

Sont donc successivement présentées :

- la représentation en terme de schémas des expressions et autres concepts mathématiques (fonction, équation, etc...),
- la représentation d'un modèle et des classes de modèle,
- la représentation d'un processus et des classes de processus,
- la représentation d'un système.

Une organisation hiérarchique de la base de connaissance pour EDORA est proposée dans ce formalisme de représentation notamment pour :

- la branche symbolique (objet-formel). Les équations différentielles se raccrochant à cette branche, on en ébauchera la classification,
- la branche modélisation (objet-modele),
- la branche biologique (objet-biologique).

II - PREAMBULE

Quelques rappels nous semblaient utiles pour préciser ce que nous entendons par modèle et modélisation. Le biologiste et le mathématicien utilisent un vocabulaire commun : système équation, variable, état, etc... Pourtant le même mot ne représente pas toujours exactement la même notion pour l'un et l'autre. Parfois elle est même différente. Tandis que pour le biologiste "équation" est par exemple associé à équation chimique, pour le mathématicien, il s'agira d'une équation différentielle ou algébrique, etc...

A l'intérieur même des mathématiques, la notion peut varier selon le contexte de la spécialité en question. Ainsi "modèle", "variable" aura un sens différent selon qu'il est employé par un spécialiste de l'automatique ou de la logique. L'automaticien sous-entend dans variable : état, entrée sortie [3], pour le logicien il s'agira d'un symbole de variables du langage formel, pour d'autres une "variable" est l'argument d'une fonction etc...

Ce point est illustré en précisant rapidement la notion de "modèle" et d'identification au sens de l'automatique, discipline donnée avec son propre formalisme.

Nous choisissons de préciser la notion de modèle dans ce cadre parce qu'elle est similaire à celle définie dans le shéma "modèle" pour EDORA. De plus, elle est tout à fait précise.

Il s'agit d'une représentation mathématique d'un système réel caractérisé par différentes sortes de grandeurs physiques reliées entre elle par une transformation. En automatique la notion de modèle est liée à celle de système dynamique.

Un système dynamique peut être schématiquement conçu comme une transformation

de grandeurs particulières (les entrées) en d'autres grandeurs (les sorties) distinguées des premières parce qu'elles donnent les caractéristiques de ce qui est produit par le système.

En automatique, cette représentation mathématique présente l'avantage d'avoir une définition rigoureuse en terme de description interne ou de description externe d'un système.

Rappels :

- la description externe d'un système est une relation d'entrée-sortie formalisée en terme mathématique par une fonction f

$y(t) = f(u(.),t)$ où t désigne le temps
u désigne l'entrée
y désigne la sortie

- la description interne par vecteur d'état est formalisée en terme mathématique par deux fonctions :

$x(t) = f(t, t_0, x_0, u)$ fonction de transition d'état
$s(t) = g(x(t), u(t), t)$ fonction de sortie

Dans le cas des systèmes différentiels, la fonction d'état est définie indirectement par une équation différentielle équation d'état

$$\dot{x} = h(x, u(t), t).$$

Il existe deux types de représentation schématique des systèmes pour l'automatique:

diagrammes fonctionnels
graphes de fluence.

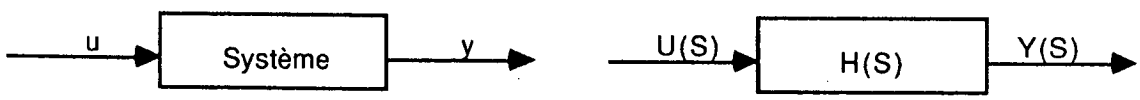
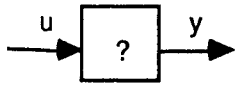


diagramme fonctionnel

graphe de fluence

"L'identification est une notion très large qu'on peut définir comme l'induction de modèles mathématiques à partir d'expériences et de connaissances à priori sur les systèmes". Telle est la définition que donnent les automaticiens de l'identification.

Pour l'automaticien le problème de l'identification est le suivant : étant donné une ou plusieurs expériences (soit des couples d'entrée-sortie {u(T), y(T)}, plus des connaissances à priori sur la structure du système considéré, en trouver une description mathématique (modèle mathématique), soit schématiquement



C'est le sens qui est donné au terme de modélisation dans un contexte comme celui d'EDORA, système à base de connaissances d'aide à la modélisation en biologie. Le terme d'identification est réservé au problème suivant: estimer les valeurs des paramètres p permettant d'ajuster les données, lorsque des connaissances a priori du système réel ont permis de retenir une classe de fonction f. En d'autres termes, il s'agit d'identifier les paramètres p d'un modèle

$$\dot{x} = f(x,u,t\ ;\ p)$$
$$y = g(x,u,t\ ;\ p)$$

où f est donnée

Cette différence peut tenir au fait que pour l'automatique classique, une connaissance a priori sur la structure du système ne peut être traduite qu'en termes de connaissance mathématique, en conséquence une connaissance a priori du système équivaut à un connaissance de la fonction f. Trouver le modèle mathématique en ce sens devient alors effectivement synonyme d'identifier les paramètres.

Pour EDORA, système à base de connaissances centré-objet, le cas est différent. Nous pouvons avoir une connaissance a priori du système autre que mathématique. A partir d'expériences et/ou de connaissances à priori biologiques, une classe de modèle mathématique peut être retenue ou rejetée.

III - NOYAU MINIMAL DE CALCUL FORMEL

3.1 - EXPRESSIONS MATHEMATIQUES

3.1.1 Motivations de la représentation d'une expression en tant qu'objet SHIRKA

Le but du projet EDORA est de fournir aux biologistes un environnement informatique capable de les aider au cour d'un processus de modélisation. Les différentes étapes de ce processus, choix ou construction de modèle, analyse qualitative du modèle, simulation, identification, nécessitent certains outils de manipulation symbolique en particulier certains outils du calcul formel.

L'objectif est de donner à un système à base de connaissances centrées-objet comme EDORA les moyens d'accès à de tels outils.

L'évolution actuelle du projet fait apparaître des besoins dans les domaines suivants:

. l'évaluation des expressions arithmétiques,
. la dérivation formelle des expressions,
. la simplification formelle,
. la "manipulation formelle" des modèles (décomposition-synthèse-reparamétrage)
. la reconnaissance et l'élaboration de modèles.

Parallèlement, les expériences et les réflexions menées dans le cadre du projet ont bien montré la nécessité de décrire dans la base de connaissance la sémantique attachée aux objets mathématiques de base manipulés.

Considérons par exemple, le cas des variables d'un modèle dynamique. Ces variables contiennent une sémantique très forte: au niveau modélisation on peut distinguer paramètres, variables d'état, variables d'entrée-sortie, temps. Cette sémantique est enrichie lorsque l'on tient compte du domaine d'application. Dans notre cas par exemple, chaque variable représente une grandeur morphologique possédant ses propres caractéristiques (unité, dimension, intervalle de valeurs...).

Il est donc nécessaire d'inclure dans la base les différentes propriétés et caractéristiques d'une variable sous la forme d'un objet SHIRKA. A cet effet, la méthode la plus cohérente et la plus simple consiste à décrire les variables et donc les expressions en tant qu'objets SHIRKA.

De la même manière, il est nécessaire de conserver les propriétés mathématiques d'un objet mathématique (par exemple, expression symbolique représentant le second membre d'un système d'équations différentielles de forme non quadratique, matrice semi définie positive) qui peuvent déterminer les traitements à appliquer sur ces objets (méthodes d'intégration, d'inversion, etc...)

Chacun des objets mathématiques étant décrit sous forme d'objets, il faut définir les mécanismes permettant leur manipulation formelle. Dans cette optique, on peut faire les remarques suivantes :

- Dans la mesure ou l'utilisateur a accès à des logiciels de calcul formel comme Macsyma ou Reduce, et à des machines suffisament puissantes pour les faire fonctionner, ces logiciels sont susceptibles de couvrir la majeure partie des besoins de calcul formel du système EDORA.

- Il est néanmoins difficile de préjuger de cette situation idéale et il convient donc, selon les besoins et pour les calculs les plus simples, d'écrire un ensemble plus réduit d'algorithmes en tenant compte éventuellement des spécificités du domaine bilologique.

- Certaines opérations faisant intervenir des manipulations symboliques, notamment la reconnaissance automatique de modèles mathématiques, sont intimement liées à des problèmes de représentation des connaissances pour lesquels SHIRKA peut apporter des solutions locales et originales.

En matière de calcul formel, le système EDORA doit donc pouvoir disposer d'interfaces de traduction lui permettant, à partir d'une représentation-objet des expressions arithmétiques, de piloter des logiciels déjà existants (Macsyma [10], Reduce, Fortran) ou à réaliser (Le lisp [2] et Shirka).

3.1.2 Représentation des expressions arithmétiques

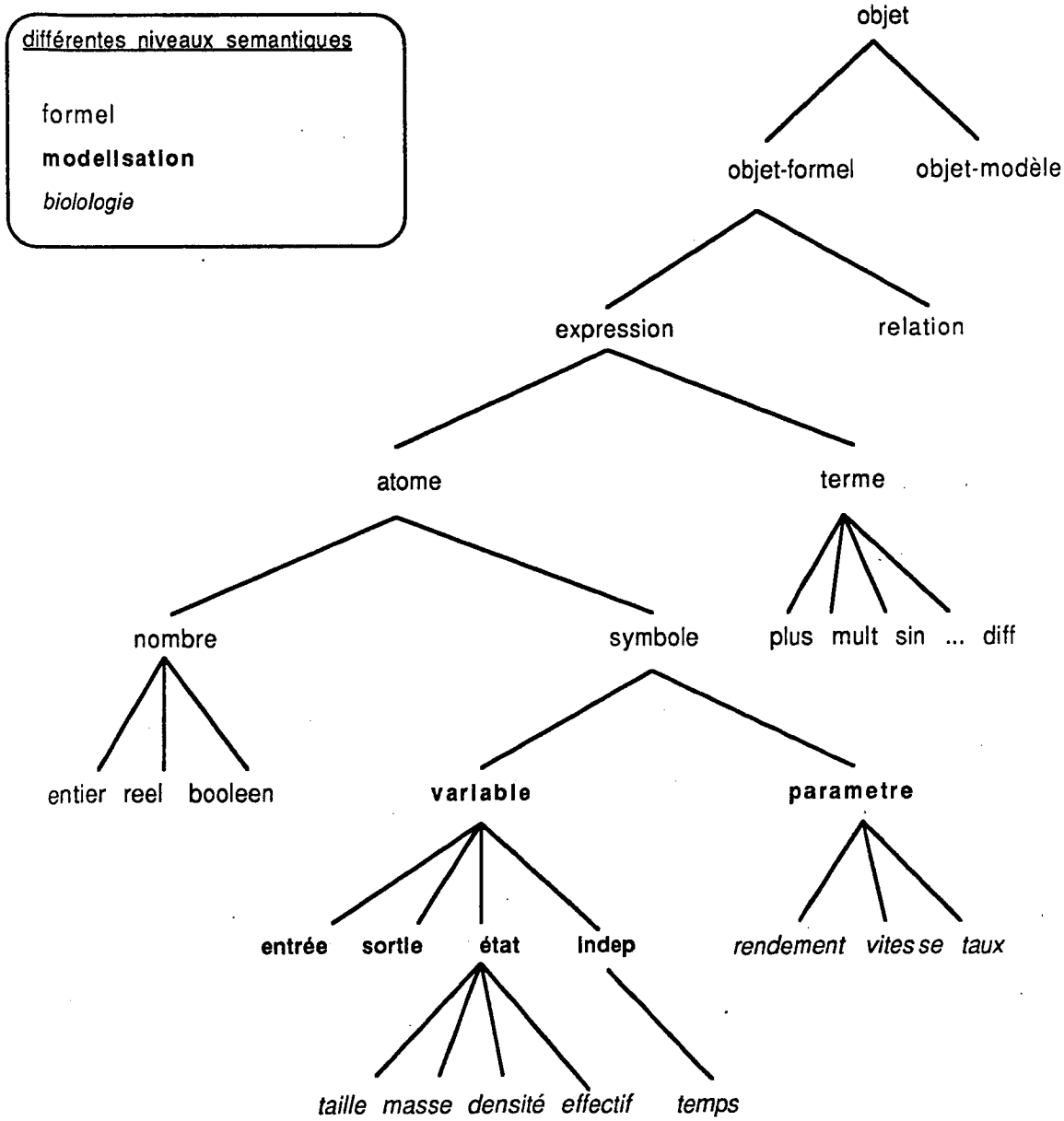
Une expression peut se représenter sous forme d'un arbre dont les nœuds sont les opérateurs (opérateurs arithmétiques ou fonctions) et dont les feuilles sont des nombres ou des symboles.

On distingue les expressions simples = atome et les expressions composées = terme

Un terme est caractérisé par un nœud et des arguments. Ses arguments peuvent être n'importe quelle expression, c'est-à-dire un atome ou un terme. A un terme, on peut également ajouter des attributs supplémentaires, par exemple décrivant :

- les méthodes d'évaluation et de dérivation propres de ce terme,
- les méthodes de traduction en Fortran ou Macsyma,
- diverses propriétés de l'opérateur peuvent être utilisées par exemple pour la reconnaissance automatique des modèles.

3.1.3 Hiérarchie des expressions orientée vers la dynamique des populations.



3.1.4 Exemples

f(x) sera représenté par l'instance

```
(terme 1
   est-un      =     terme
   nœud        =     $ f
   arguments =       $ x)
```

L'expression ax(1-x/k) sera représentée par

```
($E1
     est-un      =     mult
     arguments   =     $a,$x,$E2)

($a
     est-un      =     symbole)

($x
     est-un      =     symbole)

($E2
     est-un      =     moins
     arguments   =     1,$E3)

($E3
     est-un      =     div
     arguments   =     $x,$k)

($k
     est-un      =     symbole)
```

3.1.5 Manipulations

Ce mode de représentation des expressions permet de répondre aux objectifs souhaités :

- Des commandes permettant l'évaluation numérique, la dérivation, la simplification ont été intégrées au système. Elles sont définies comme des méthodes décrites par un schéma de classe, et font appel à des fonctions Lisp adéquates. Ces fonctions Lisp d'évaluation, simplification, dérivation ont pour arguments les expressions représentées sous la forme d'objet décrite ci-dessus.

- Il offre une possibilité d'extension aux autres notions mathématiques indispensables pour la modélisation et la simulation : équation différentielle, fonction.

Sont présentés d'abord quelques exemples de représentation de notions mathématiques à l'aide de schémas puis l'organisation générale de la hiérarchie des objets mathématiques.

3.2 EXEMPLES D'AUTRES OBJETS MATHEMATIQUES OU SYMBOLIQUES

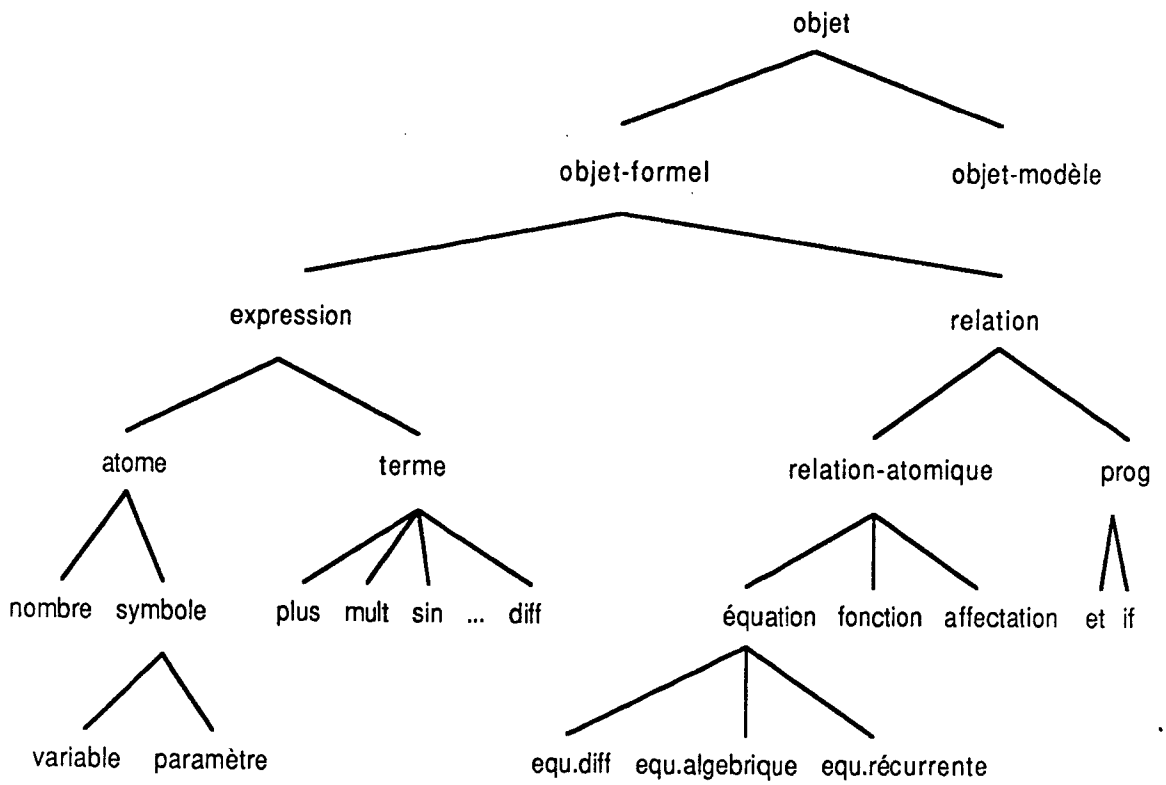
NOTION	**SCHEMAS**
1 équation y = ax	(équation#1 est-un = équation arguments = (y mult)) (mult 1 est-un = mult arguments = (a x))
2 fonction f(x) : =ax	(fonction 1 est-un = fonction arguments = (terme1 mult1)) (terme 1 est-un = terme nœud = f arguments = (x))
3 affectation y : ax	(affectation 1 est-un = affectation arguments = (y mult1))
4 prog x : 1 y = ax z = ay	(prog 1 est-un = prog local = (affectation 1) arguments = (équation1 équation2)) (affectation 1 est-un = affectation arguments = (x 1)) (équation 1 est-un = équation arguments = y mult1) (équation 2 est-un = équation arguments = z mult2)

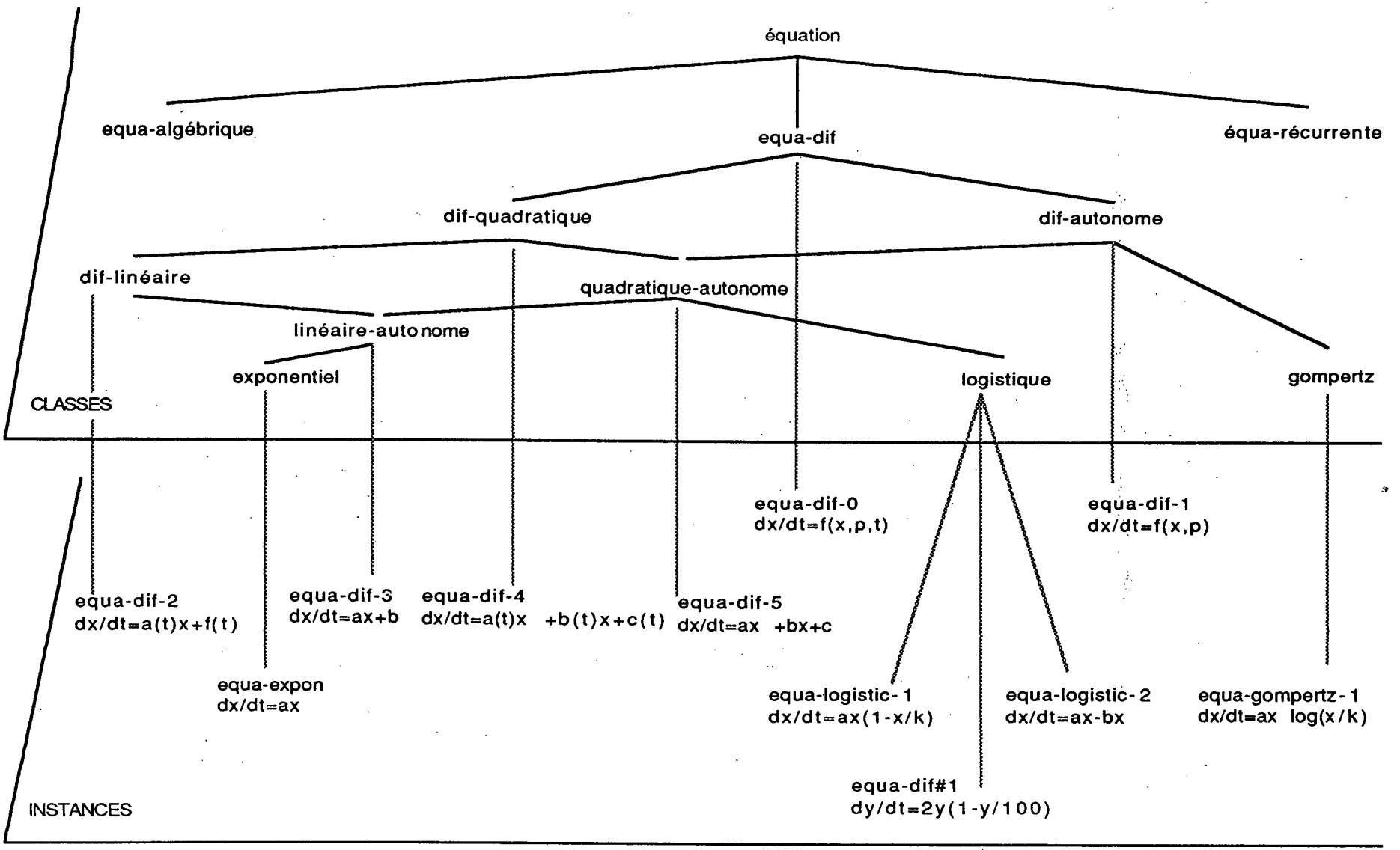
```
5   équation différentielle
    dx/dt = ax      (équation-différentielle 1
                            est-un          =   équation-différentielle
                            membre-gauche   =   diff1
                            membre-droite   =   mult1)

                    (diff 1
                            est-un          =   diff
                            arguments       =   $x$t)

                    (mult 1
                            est-un          =   mult
                            arguments       =   $a$x)
```

ORGANISATION GENERALE DE LA HIERARCHIE DES OBJETS MATHEMATIQUE





Toutes les instances soulignées sont des instances prototypiques

IV - REPRESENTATION D'UN MODELE

4.1 Variables d'un modèle

4.1.1 Spécialisations de l'objet "symbole" : "variable", "paramètre"

- Sur le plan formel ,dans l'expression ax(1-x/k) on ne peut distinguer que deux types d'atomes : les symboles : a, x, k qui peuvent être affectés, par opposition aux "constantes" : 1 (cf. 3.2).

Comme on l'a rappelé, un modèle est la représentation mathématique d'un système biologique. Il est défini par :

(1) $x(t) = f(x(t), t, u(t); p)$
$y(t) = g(x(t), u(t); p)$
$x(t_0) = 0$

où (2)

$x(t) \in R^m$ est le vecteur d'état
$p \in R^m$ représente les m coefficients inconnus
$t \in [0,T]$ est le temps
$y(t) \in R^p$ est la sortie

Les symboles intervenant dans les expressions mathématiques du modèle véhiculent une sémantique propre liée à la modélisation (2). C'est pourquoi on distingue deux sortes de symbole : les paramètres et les variables. Les paramètres sont les objets à identifier (ils doivent être affectés avant l'appel à la simulation et sont constants au cours de la simulation). Les variables sont spécialisées en état, entrée, sortie, variable-indépendante.

4.1.2 Spécialisations intégrant la sémantique biologique

Lorsqu'on veut modéliser une situation donnée ,les paramètres, les variables d'état du modèle représentent des grandeurs physiques précises : lorsqu'on s'intérese à une population la variable d'état représentera par exemple un effectif. Pour un organisme, il s'agira plutôt de grandeurs morphologiques : taille, poids, etc... Un paramètre représentera une vitesse, un rendement de réaction, un taux de croissance, etc... Les classes état et paramètres peuvent donc être spécialisées en tenant compte de la sémantique biologique.

4.1.3 Représentation en terme d'objet des variables d'un modèle

L'organisation hiérarchique des objets symboliques proposée en 3.3 permet d'intégrer ces informations. Certaines caractéristiques peuvent être inférées à partir de celles-ci : domaine de variation, unité, dimension, etc... La valeur de certains attributs est prédéfinie au niveau d'une classe grâce à la facette $valeur. Par exemple si la variable x est une instance de taille, la valeur "L" de son attribut dimension est héritée de cette classe.

La représentation centrée objet permet de réunir sur un même objet ces diverses connaissances.

(variable		
sorte-de	=	symbole
signification	$un	chaine
domaine-définition	$un	intervalle
unité	$un	chaine
dimension	$un	symbole)
(paramètre		
sorte-de	=	symbole
signification	$un	chaine
domaine-variation	$un	intervalle
unité	$un	chaine
dimension	$un	symbole)
(état		
sorte-de	=	variable)

et par exemple

(taille		
sorte-de	=	état
signification	$valeur	"taille"
domaine-définition	$valeur	R^+
unité	$défaut	"cm"
dimension	$valeur	L)

4.2 Objet-modèle

La définition de l'objet-modèle est alors la suivante :

(modèle		
est-un	=	sch-modèle
sorte-de	=	objet
var-indep	$un	var-indep
état	$liste-de	état
paramètre	$liste-de	paramètre
entrée	$liste-de	entrée
sortie	$liste-de	sortie
cond-ini	$un	atome
équation	$liste-de	equa-dif
prototype	$un	modèle
etc...		)

4.3 Définition d'un modèle en tant que classe

Le modèle dx/dt = 1.1x(1 - x/100) ; x(0) = 10 est une instance particulière du modèle logistique dont la structure est définie par dx/dt = ax(1 - x/k) ; x(0) = xo. Pour caractériser cette classe on définit un représentant privilégié dit instance prototypique. Ces instances prototypiques permettent :

. de savoir si un modèle introduit par l'utilisateur correspond à un modèle de la base,
. de créer une instance d'une classe de modèle par "duplication" du prototype,
. de pouvoir donner un exemple de modèle pour chaque classe.

Ce mode de représentation des classes par un représentant canonique n'a de sens que si on définit une relation d'équivalence entre les modèles, c'est-à-dire si on possède un mécanisme de mise en correspondance (ce qui implique une étape de simplification formelle). Un algorithme de reconnaissance a donc été implémenté. Il utilise la forme pseudo-canonique d'une expression obtenue à la sortie du simplificateur. Les spécifications du simplificateur ont été fixées en liaison étroite avec par exemple ce type de fonctionnalités. (cf. 3.5).

4.4 Exemple : schéma de classe modèle logistique et instance prototypique

```
(modèle-logistique
     (est-un  (= sch-modèle))
     (sorte-de  (= modèle-différentiel))))

(modèle-logistique-forme1
   • (est-un  (= sch-modèle))
     (sorte-de  (=modèle-logistique))
     (prototype  ($défaut  (prototype-logistic-1))))
```

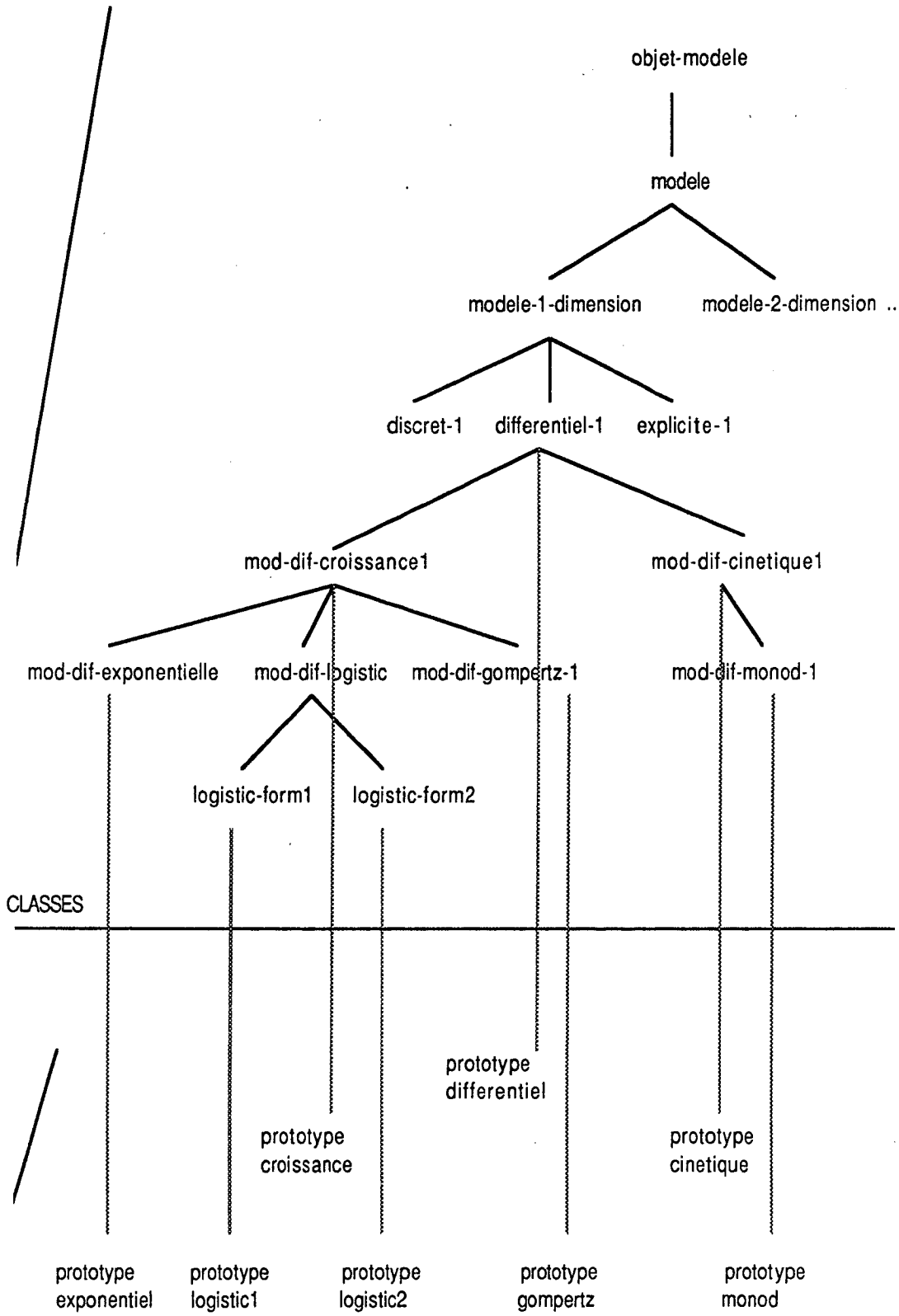
Instance prototypique définissant la classe modèle-logistique-forme1

```
(prototype-logistic-1
     (est-un  (= modèle-logsitique-forme1))
     (état  (= ($%%x)))
     (paramètres  (= $%%a $%%k)))
     (équations  (= equ-logistic-1)))))
```

equ-logistic-1 est une instance prédéfinie dans la base de connaissance des équations différentielle (cf. classification des équations différentielles).

La présentation ci-dessous de la branche objet-model n'est qu'une ébauche. Pour plus de précisions on pourra se reporter à des références de spécialistes [6], [8], [13].

4.5 Organisation hiérarchique des modèles



4.6 Exemple d'instance de logistic

soit le modèle logistique défini par
dx/dt= 1.1*x*(1-x/100.)
x(o) = 50
où x représente la taille d'un rat musqué mesurée en cm étudiée pendant 36 mois
;;unité choisie...
t représente le temps, domaine d'étude [0+inf]

```
(exemple-logistique-forme1
        est-un          = modèle-logistique-forme1
        état            = $taille-rat-musqué
        paramètres      = (1. , 1100.)
        équations       = equ-dif1
        cond-ini        = 50)

(equ-dif1
        est-un          = equation-dif
        membre-gauche   = diff1
        membre-droit    = expr1)
```

membre de gauche de l'equation différentielle equ-dif1

```
(diff1
        est-un          = diff
        arguments       = $taille-rat-musqué $t)
```

membre de droite de equ-dif1

```
(expr1
        est-un          = mult
        arguments       = (1.1, $taille-rat-musqué))

(expr2
        est-un          = moins
        arguments       = ( 1 $expr3))

(expr3
        est-un          = div
        arguments       = ($rats-musqués 100. ))
```

variables mathématiques:

($taille-rat-musqué
est-un = taille
domaine de définition = [0 60]
unité = cm
dimension = $L
mesures = longitudinales)

($t
est-un = temps
signifiant = temps
domaine de définition = R+
dimension = $T
unité = mois)

On a souligné les valeurs d'attributs héritées de la classe auxquelles appartiennent les instances.

V - FORMALISME BIOLOGIQUE : PROCESSUS

5.1 Formalisme pseudo-chimique

Il est nécéssaire d'introduire un formalisme plus proche du "langage naturel" du biologiste lui permettant de décrire la situation étudiée. Dans de nombreux domaines, des représentations schématiques ont été utilisées pour décrire certaines hypothèses sur la structure ou le fonctionnement du système : diagrammes "boîtes-fléches", "bond graphs", diagrammes de Forrester, etc... Après avoir remarqué la parenté entre les modèles de la cinétique chimique et ceux de la dynamique des populations, la "représentation pseudo-chimique", voisine de celle utilisée pour représenter les équations chimiques a été retenue [4] [3].

Le système est défini à l'aide de deux sortes d'objets : les espèces $X_1, X_2, ..., X_q$ qui interagissent et les réactions élémentaires décrivant leurs interactions. L'évolution du système est ainsi décrite par une liste de p réactions :

$$(R)\quad n_{i1}X_1 + n_{i2}X_2 + ... + n_{iq}X_q \xrightarrow{a_i} m_{i1}X_1 + m_{i2}X_2 + ... + m_{iq}X_q \quad (1 \le i \le p)$$

Cette notation s'interprète comme l'intéraction des termes en partie gauche (en proportion n_{ij}) pour produire les termes en partie droite (en proportion m_{ij}). Les a_i symbolisent les constantes de vitesse de réaction. Cette formulation est fondée sur des hypothèses dérivées des lois du type "action de masse".

La procédure de traduction formalisée dès 1961 par Garfinkel [4] permet d'associer au système (R) le système d'équations différentielles :

$$dX_j/dt = \sum_{i=1}^{P} a_i(m_{ij} - n_{ij}) \prod_{k=1}^{q} (X_k)^{n_{ik}} \qquad (1 \leq j \leq q)$$

5.2 Notion de processus

Un "processus" est la représentation dans le formalisme pseudo-chimique d'un système biologique. Il est défini par :

$$(1)\quad n_{i1}X_1 + \ldots + n_{iq}X_q \xrightarrow{a_i} m_{i1}X_1 + \ldots + m_{iq}X_q \qquad 1 \leq i \leq P$$

où

(2)
- $X \subset E^q$ vecteur d'entité
- $n_{ij} \subset R$ représente des proportions inconnues
- a_i la constante de réaction

Ce formalisme permet de représenter les processus de base mis en oeuvre en dynamique des populations (consommation de substrat, vieillissement, mortalité, renouvellement, compétition, etc...). Il offre une possibilité de construction de nouveaux processus par combinaison de ces processus de base.

5.2.1 Exemples de processus

La croissance d'une population X par consommation d'un nutriment S peut être représentée par le processus 1

$$(R1)\quad X+S \xrightarrow{a} (1+R)X$$

La production du substrat S par dégradation de la population X est définie par le processus 2

$$(R2)\quad X \xrightarrow{b} pS$$

où
- X est la population
- S le rendement
- b la constante
- p proportion de produit

On peut créer un nouveau processus par combinaison des processus 1 et 2 .Il sera défini à l'aide de R3 = R1 $\wedge$ R2 :

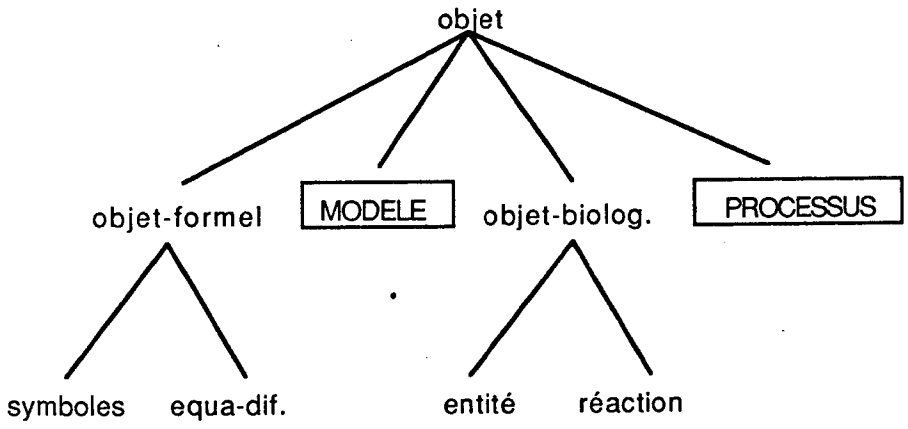
$$(R3)\quad \begin{cases} X + S \xrightarrow{a} (1 + R)X \\ X \xrightarrow{b} pS. \end{cases}$$

5.3 Intégration des objets biologiques dans EDORA : représentation d'un processus

La description d'un système biologique dans un formalisme donné nécessite de représenter en terme d'objet SHIRKA "le langage de description" propre au formalisme.

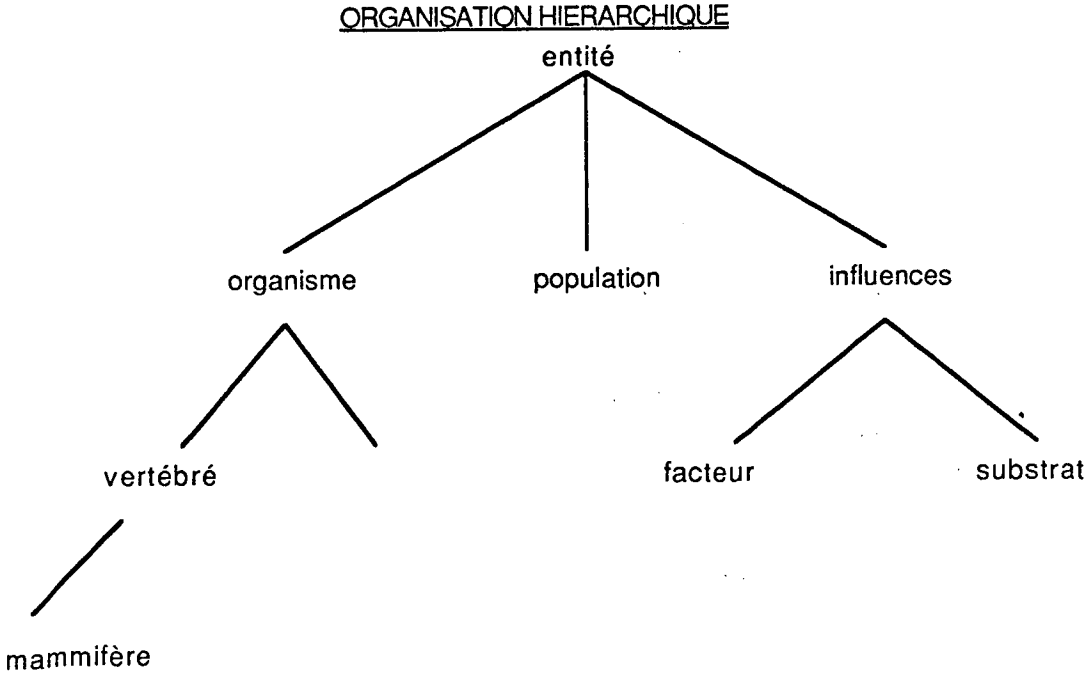
Pour représenter un "modèle" en terme d'objet Shirka, il a fallu définir un certain nombre d'objets mathématiques : variables, équation-différentielle.

Pour représenter "un processus" il faut définir les objets spécifiques à ce formalisme : "entité", "réaction".



5.3.1 Entité

Les espèces en présence sont des instances de la classe "entité". On n'utilise parsle terme de variable pour éviter la confusion avec l'objet mathématique "variable" défini en 4.1. Conformément à la sémantique biologique, différentes spécialisations sont proposées : les espèces étudiées peuvent être des instances des classes : "organe", "organisme", "population", ou des éléments influents l'évolution : "facteur", "substrat".



La classe "entité" possède un attribut particulier : "variables" permettant de préciser les grandeurs quantifiables caractéristiques de la classe : variables morphométriques décrivant un individu, quantité de produit (pour un substrat), effectif pour une population etc...

Exemples

Variables biologiques

```
(X
   est-un        = entité
   signification = "cellule"
   variables     = $x)

(Y
   est-un        = mammifère
   signification = "rat-musqué"
   variables     = $x $y
   habitat       = "étang"
   etc ...                  )
```

Variables mathématiques

```
($x
   est-un        = variable
   signification = "diamètre cellulaire"
   domaine       = R+
   unité         = µ)

($x
   est-un        = poids)

($y
   est-un        = taille
   signification = "longueur-de-patte")
```

5.3.2 Réaction

La classe réaction est caractérisée par ses attributs membre de gauche, membre de droite, constante de réaction. Chaque membre a pour valeur une liste de termes chimiques élémentaires.

Exemple

La réaction $X + S \xrightarrow{a} (1 + R)X$ sera une instance de réaction définie par :

```
(1)  (R
        est-un       = réaction
        mb-gche      = chim1 chim2
        mb-dte       = chim3
        constante    = a)

     (chim1
        est-un       = réaction
        variables    = X
        coefficient  = 1)

     (chim2
        est-un       = terme-chimique
        variables    = S
        coefficient  = 1)

     (chim3
        est-un       = terme-chimique
        variables    = X
        coefficient  = plus1)
```

où X et S sont des instances d'entité
a et R sont des instances de paramètres
plus1 est une instance d'expression "plus" représentant l'expression (1+R).

5.3.3 Classe processus

La classe processus est définie par le schéma :

```
(processus
   sorte-de      =          objet
   variables     $liste-de  entité
   réaction      $liste-de  réaction
   prototype     $un        processus).
```

Exemple d'instances de processus

La croissance d'Escherichia Coli en milieu liquide complexe sera représentée par le processus P.

```
(P
   est-un        =          croissance-consommation-substrat
   variables     =          (X  S)
   réactions     =          R)     ( définie en 4.3.2 -(1) )

(X
   est-un        =          population
   signification =          "Escherichia-Coli"
   variables     =          $d)

($d
   est-un        =          densité
   signification =          "densité optique"
   unité         =          "D.O")

(S
   est-un        =          substrat
   signification =          "milieu  1").
```

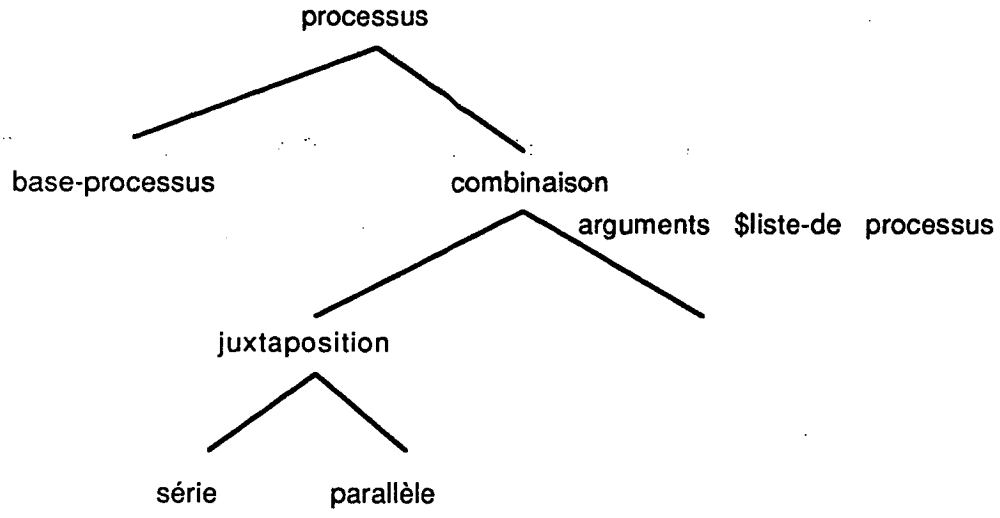
5.4 Algorithme de traduction

Un algorithme de traduction réalisé en Le Lisp permet de passer de la représentation biologique à la représentation mathématique. A partir d'un "processus" représenté sous forme de schémas, il génère l'instance de modèle correspndante (cf. 4.1). Cet algorithme utilise les facilités introduites par l'intégration dans Edora du "calcul formel" minimal défini en III.

5.5 Organisation hiérarchique des processus

Afin de pouvoir construire de nouveaux processus, on doit disposer d'une base de processus et de mécanismes primitifs de combinaison de ces processus de base. L'organisation proposée correspond à une "construction incrémentale de processus".

Il s'agit donc de constituer une base-de processus dont l'organisation hiérarchique est satisfaisante du point de vue des biologiques (cf. hiérarchie page). Comme pour les modèles, les processus de cette base sont définis à l'aide d'instances prototypiques



La combinaison la plus simple de processus de cette base est la juxtaposition notée "."
Ainsi si

```
(P1                              (P2
   est-un     = processus            est-un     = processus
   variables  = (A B)                variables  = (C D)
   réactions  = A -> B)              réactions  = C -> D)
```

et si P3 = P1 . P2 alors le résultat P3 sera défini par

```
(P3
    est-un     =  juxtaposition
   variables =  A  B  C  D
   réactions =   ( A  -> B ;  C -> D))
```

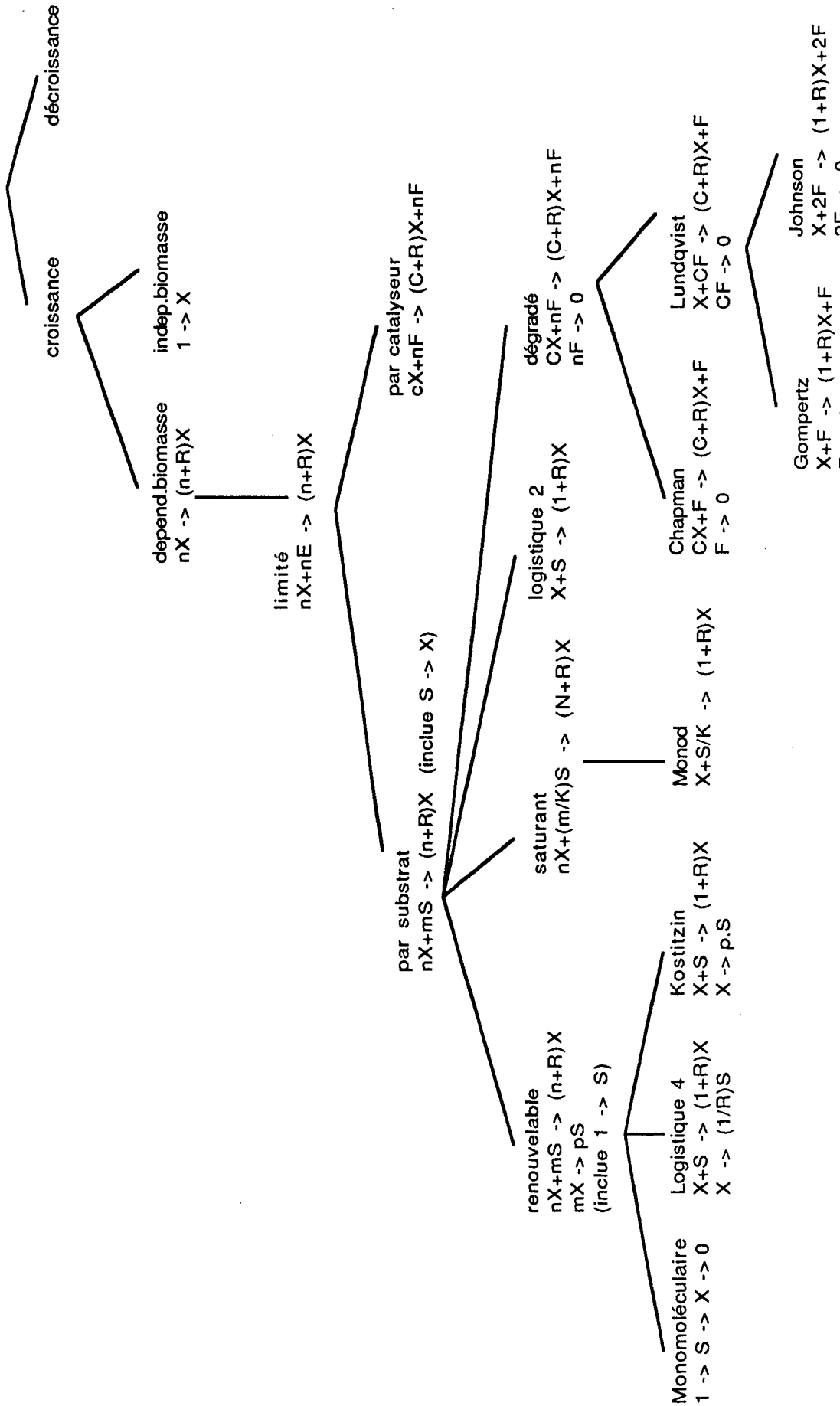
Certaines spécialisations peuvent être prédéfinies : série (A -> B -> C), parallèles distinguées en

jumelles A + B -> C + D
A + B -> C' + D'

ou compétitives A + B -> C + D
A' + B -> C' + D'

etc...)

Approche d'organisation hiérarchique d'une base de processus



VI - SYSTEME

6.1 Définition de la classe "système"

La notion de système dynamique au sens de l'automatique ne contient qu'une sémantique quantifiable : il est caractérisé par des grandeurs physiques reliées entre elles par une transformation traductible uniquement en loi mathématique.

Lorsqu'on étudie un phénomène biologique, il est caractérisé par des espèces biologiques reliées entre elles par des lois biologiques.
Différentes interprétations du système étudié peuvent être données, selon le point de vue retenu.

L'objectif est de réunir sur le même objet toute la sémantique, tant mathématique que biologique, l'intérêt d'une représentation de la connaissance centrée objet est d'offrir cette possibilité.
Ainsi l'objet "système" résume toute l'information. Il est défini par le schéma de classe :

```
(système
   sorte-de      =        objet
   modèle        $un      sch-modèle
   processus     $un      sch-processus
   données       $un      donnée
   situation     $un      situation).
```

6.2 Exemple

Le phénomène étudié est la croissance du rat musqué [A. Pavé]. L'instance correspondant à ce problème est :

```
(S1
   est-un        =        système
   modèle        =        modèle1
   processus     =        processus1
   données       =        données1
   situation     =        situation1)
```

- La description mathématique est fournie par

```
(modèle1
   est-un        =        gompertz
   état          =        x  y
   paramètres    =        a  b
   cond-ini      =        x0
   équations     =        equa-dif1   equa-dif2)
```

où

equa-dif1 représente l'équation différentielle dx/dt = -ax
equa-dif2 représente l'équation différentielle dy/dt = bxy
x représente la biomasse du rat

y caractérise l'hormone de croissance
a et b sont des paramètres de dimension respectives T^{-1} et (unité (x))$^{-1}.T^{-1}$

- L'interprétation biologique est fournie par

(processus1
est-un = croissance
variables = X F
réactions = réaction1 réaction2)

où

réaction1 représente l'équation pseudo-chimique $X + F \xrightarrow{b} (1+R)X+F$

réaction2 représente l'équation chimique $F \xrightarrow{a} 0$
X représente le rat-musqué
F représente le facteur hormonal de croissance.

Ainsi processus1 représente une auto-production de X catalysée par un facteur F se dégradant indépendamment de X.

- Les données sont définies par une instance

(données1
est-un = donnée
chronique = chronique1
forme = forme1)

chronique1 est la suite de mesures n_i aux instants t_i

forme1 est une instance de forme par exemple

(forme1
est-un = monotone-croissante)

(cf. hiérarchie de la classe forme).

- La situation est décrite par une instance de situation, par exemple :

(situation1
est-un = situation
variables = X
loi = croissance)

Différents types de situation peuvent être rencontrées. Il peut s'agir de l'évolution d'une variable attachée à un individu, de l'évolution d'un individu (mortalité d'un animal), ou de l'interaction entre différentes populations, par exemple un phénomène de prédation sera représenté par:

(situation2
est-un = situation
variables = lièvres lynx
loi = prédation)

L'exemple du rat musqué est en fait un exemple de problème résolu. Cette situation

biologique ayant déja été étudiée [14], l'interprétation biologique de la situation et le modèle mathématique approprié sont connus.

En géneral, ce n'est pas le cas. Il s'agit d'inférer la valeur des attributs modèle ou processus à partir des informations incomplètes auxquelles on a accès. La forme des données, la description de la situation peuvent permettre de proposer plusieurs modèles ou processus. Le système doit conduire ce processus de modélisation, guidé par les informations et les choix que le biologiste est capable d'exprimer.

VII - CHOIX OU CONSTRUCTION DE MODELES

L'intégration dans le système des objets symboliques et biologiques permet au logiciel de piloter les différentes étapes d'un processus de modélisation en tenant compte des diverses informations : données, situations, modèle, processus. Les étapes de choix et de validation du modèle sont abordées ci-dessous.

7.1 Choix d'une classe de modèles

Deux niveaux peuvent être distingués :

1 - Reconnaissance

Il s'agit de reconnaître si le modèle, le processus ou la situation introduite par l'utilisateur existe déjà dans la base de connaissance. Si tel est le cas, le modèle correspondant est fourni avec toute l'information qui y est attachée. Par exemple, croissance de rats musqués, croissance de vertébrés, etc... sont des situations qui peuvent être reconnues dans la base. Si le biologiste entre le modèle : ax(1-x/k), le modèle est également reconnu en tant que logistique.

2 - Classification

Si plusieurs informations sont disponibles, il s'agit de gérer leur cohérence et de proposer une liste de modèles et de processus possibles.

7.2 Validation

L'intégration du formalisme biologique est importante pour cette étape qui soulève des questions d'identifiabilité, de discernabilité, etc...

Les travaux d'Eric Walter sur ces notions d'identifiabilité et de discernabilité structurelle posent les problèmes auxquels on est régulièrement confronté lors de la modélisation en biologie [17]. Il propose pour tester les propriétés structurelles d'un modèle une solution faisant appel au calcul formel sur des polynomes [9].

- s'il existe des modèles de structère différente ayant le même comportement ("modèles structurellement non discernables"),

- s'il existe plusieurs modèles de même structure ayant le même comportement ("modèles structurellement non globalement identifiable"),

- alors on souhaite
  1) les trouver tous
  2) éliminer ceux qui ne peuvent être validés

Après une étude théorique de ces questions, seule la prise en compte des aspects biologiques peut être discriminante.

Parmi les modèles retenus, des modèles de structure différente peuvent avoir une aussi bonne valeur descriptive des données expérimentales. Par contre les connaissances biologiques peuvent être déterminantes pour le choix d'une classe de modèle (cf. modèle de gompertz pour la croissance des rats musqués [14].)

Lorsque plusieurs valeurs des paramètres sont proposées, les valeurs aberrantes (quant à leur signification biologique) peuvent être éliminées : rejet de valeur négative pour une concentration, etc...)

7.3 Construction

L'organisation de la base de connaissance proposée, permet de construire de nouveaux modèles ou de nouveaux processus à partir des bases de modèles ou de processus. Si aucun modèle n'est satisfaisant, un nouveau modèle pourra être proposé par combinaison des différents processus de base. Une approche similaire a été proposé par [12] [18]...

VIII - CONCLUSION

L'élaboration et l'implémentation de la base de connaissance présentée ci-dessus constitue un premier pas vers un environnement homogène utilisant la connaissance codée sur ses objets pour assister ou guider le biologiste dans sa démarche de modèlisation.

La structuration proposée de la base fait clairement ressortir les avantages d'une représentation centrée objet pour un système expert d'aide à la modèlisation comme EDORA

- représentation unifiée des données, des méthodes et des objets manipulés par ces méthodes (combinaison prodédural/déclaratif).
- intégration de la sémantique biologique sur des objets qui se "dépouilleraient de leur signification" dans d'autres formalismes de représentation (par exemple les variables d'un modèle et le modèle lui même).
- possibilité de faire coexister des outils et compétences de disciplines aussi éloignées que calcul formel (liée a un domaine méthodologique) et biologie (liée à un domaine d'application).
- facilité d'évolution de la base de connaissances.

Il est toutefois clair que la construction d'une telle base de connaissance nécessite un effort de conception et est loin d'être achevée...

BIBLIOGRAPHIE

[1] Bobrow D.G., Winograd T., An Overview of KRL, a Knowledge Representation Language. *Cognitive Science,* 1, 1, 1977, 2-46.

[2] Chailloux J., "Le Lisp, Le manuel de référence". Rapport technique INRIA n° 27 (Juillet 1983).

[3] Faurre P., DEPEYROT M., Elements d'automatique. Dunod, Paris, 1974.

[4] Garfinkel D., Rutledge J.D., Higgins J.J., Simulation and Analysis of Biochemical Systems : I. Representation of Chemical Kinetics. *Comm. of the ACM,* 1961, 559-562.

[5] Hamrouni M.K., Etude et développement d'un système informatique d'aide à l'élaboration de modèles en biologie. Dr 3ème cycle thesis, Université Pierre et Marie Curie, Paris, 1979.

[6] Jolivet E., Introduction aux modèles mathématiques en biologie, Masson, 1983.

[7] Kreisel G., Krivine J.L., Eléments de logique mathématique. Théorie des modèles, Dunod, 1967.

[8] Lebreton J.D., Millier C., Modèles dynamiques déterministes en Biologie. Masson, Paris, 1982, 207.

[9] Lecourtier Y., Raksanyi A., The testing of structural properties through symbolic computation. (à paraître).

[10] MACSYMA. Ref. Manual, The Mahlab Group, *Lab. for Computer Science,* MIT, 1982.

[11] Minsky M., A Framework for Representing Knowledge. in "*The Psychology of Computer Vision*", Ed. Winston P.H., Mc Graw-Hill, 1975, 211-277.

[12] Nolan P.J., Mc Carthy M.A., A.I. Frame-based Simulation in Systems Dynamics. Application of Artificial Intelligence in Engineering Problems. 1st Internatinal Conference, Southampton University, U.K.,April 1986.

[13] Pave A., Contribution à la théorie et à la pratique des modèles mathématiques pour l'analyse des systèmes biologiques. Thèse de Doctorat es Sciences, Université Claude Bernard (Lyon I), 1980.

[14] Pavé A., Corman A., Bodillier-Monot, Application à l'étude de la croissance de jeunes rats musqués. Biom. Proxim. (1986), 26, 123-140.

[15] Pavé A., Rechnemann F., Computer aided modelling in biology : an artificial intelligence approach. Artificial Intelligence and Simulation, SCS Rev.

[16] Rechenmann F., SHIRKA : mécanisme d'inférence sur une base de connaissances centrée-objet. *Proceed. of "reconnaissance des formes et I.A",* Grenoble, Nov. 1985.

[17] Walter E., Identifiability of State Spaces Models. Springer Verlag, Berlin Heidelberg, New-York, 1982.

[18] Zeigler B.P., Structuring principles for multifaceted system modeling. Methodology in systems modeling and simulation. B.P. ZEIGLER and ad (eds), NORTH HOLLAND 1979.

IDENTIFICATION DE MODELES DYNAMIQUES. ASPECTS STATISTIQUES

Antoine MESSÉAN
Laboratoire de Biométrie
INRA-CRJJ
78350 JOUY-EN-JOSAS

1 INTRODUCTION

La représentation de phénomènes biologiques par un modèle mathématique répond à des objectifs divers : prévision de l'évolution du phénomène considéré, simulation numérique de situations observables ou non, comparaison de populations ou de courbes, interprétation biologique de certains paramètres du modèle (taux de croissance ou taille maximale d'une population).

Les problèmes numériques liés à l'intégration de systèmes différentiels et à l'identification des modèles ont, très longtemps, constitué le problème essentiel au détriment des considérations statistiques et de la réflexion sur le sens des modèles. L'avénement de calculateurs puissants et le développement de nombreux algorithmes d'optimisation rendent facilement accessible l'utilisation de modèles dynamiques plus ou moins complexes.

Dans ce contexte, il est maintenant possible de s'intéresser aux aspects statistiques de la modélisation : en effet, l'identification d'un modèle ne se limite pas à son ajustement à des données par la méthode des moindres carrés. La variabilité des données observées, l'interprétation des valeurs ajustées, la nature des objectifs fixés mettent en oeuvre des méthodes et des outils statistiques plus complexes mais qui sont désormais utilisables dans la majorité des cas.

L'objet de ce rapport est de faire un rapide tour d'horizon de certains des aspects statistiques que nous rencontrons dans le cadre de l'identification des modèles : construction du modèle statistique, estimation des paramètres, validation et comparaison de modèles, inférence sur les paramètres.

2 MODELISATION.

Les modèles que nous considérons ici sont des modèles de régression, c'est-à- dire qu'ils s'expriment de la façon suivante :

$$Y_i = f(x_i, \theta) + \varepsilon_i \qquad i = 1, ..., n \tag{1}$$

où (Y_i) sont les observations, f une fonction non-linéaire par rapport à θ, θ est le vecteur des paramètres de dimension p ($\theta \in \Theta \subset R^p$), (x_i) le vecteur des variables explicatives, (ε_i) des erreurs aléatoires, centrées, indépendantes et de variance $v(x_i, \theta)$.

La modélisation consiste à se donner, d'une part, la fonction f, c'est-à-dire la partie déterministe du modèle de régression, d'autre part la distribution des erreurs ε_i, c'est-à-dire

leur loi et, surtout, la forme de leur variance. Cette seconde composante est, le plus souvent, négligée par le modélisateur alors qu'elle est tout aussi importante que la première.

2.1 Modélisation de la fonction espérance f.

Sans prétendre aborder le problème de la modélisation pour lequel de nombreux ouvrages existent (Lebreton, J.-D. & Millier C., Jolivet E., par exemple), il est important de distinguer deux approches extrêmes qui correspondent, en fait, à des objectifs très différents :

- Le modèle d'interprétation.

  C'est le cas lorsque l'objectif est d'obtenir un modèle d'interprétation dont on exploitera les paramètres en tant que quantités biologiques. Généralement, la construction du modèle se fait alors directement à partir des hypothèses de fonctionnement du phénomène considéré. Les modèles de dynamique de populations régis par des équations différentielles constituent un bon exemple de modèles d'interprétation.

- Le modèle de représentation.

  Le modèle ne sert que de représentation mathématique des observations et est alors utilisé pour d'autres objectifs que l'obtention d'un modèle "fidèle". C'est le cas, par exemple, des dosages radioimmunologiques où le modèle logistique est utilisé comme représentation d'une courbe de calibration qui est ensuite "inversée" afin d'estimer des doses de substrat (Huet, 1984). Dans ce contexte, l'avantage des modèles paramétriques par rapport aux modèles non paramétriques (polynômes ou fonctions splines) est qu'ils permettent une meilleure inférence.

2.2 Modélisation de l'erreur.

Cette modélisation de l'erreur peut s'effectuer composante par composante (la loi des erreurs, leur variabilité et leur indépendance) bien que celles-ci ne soient pas toujours dissociables.

Le point essentiel à vérifier, en tout premier lieu, est qu'il s'agit vraiment d'un problème de régression. En effet, certains phénomènes peuvent prêter à confusion : par exemple, si nous observons, jour après jour dans un lot de graines, le nombre de celles qui ont germé et si nous portons ces nombres sur un graphique en fonction du temps, les allures de courbes obtenues sont sigmoidales. On pourrait alors, à première vue, penser à appliquer les techniques de la régression alors qu'en fait, il s'agit d'un problème d'ajustement de distributions et les techniques à appliquer sont de nature très différente.

L'hypothèse d'indépendance des erreurs est généralement faite alors que nous savons que dans les phénomènes de croissance, par exemple, il n'y a pas indépendance des observations effectuées sur les mêmes individus. Cette hypothèse n'est pas fondamentale mais elle peut expliquer quelquefois le comportement périodique des résidus.

La variabilité est, sans aucun doute, la composante la plus facile à modéliser et à prendre en compte, même si les problèmes d'estimation se compliquent sensiblement. S'il y a des répétitions, c'est-à-dire si Y est observé plusieurs fois (n_i fois) pour une même combinaison des variables explicatives x_i, on étudiera les variances empiriques $s_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i}(y_{ij} - y_{i.})^2$ (où $y_{i.}$ est la moyenne des y_{ij}). Sinon, on peut "estimer" la variabilité des erreurs par celle des résidus obtenus après un premier ajustement à variance constante (en considérant, par exemple, la valeur absolue ou le carré des résidus).

La modélisation de la variance peut se faire en observant les variations de la variance avec l'espérance f (ou les observations Y) ou bien par rapport aux variables explicatives. D'autre part, la nature du phénomène considéré peut aider le modélisateur à déterminer la forme de

la variance. Par exemple, il est très fréquent, dans les modèles de croissance, de rencontrer des variances qui sont de la forme $v(x_i, \theta) = \sigma^2 f(x_i, \theta)^h$ où h est une constante, supposée connue.

Comme nous l'avons déjà dit, le point essentiel est de vérifier que le problème est bien décrit par un modèle de la forme (1). Il est possible que cette forme soit obtenue après transformation : par exemple, lorsque les erreurs ε_i interviennent de façon multiplicative, une transformation logarithmique nous ramène à la forme (1). Ces transformations peuvent avoir comme conséquence de linéariser la fonction f mais cette linéarisation du modèle f ne doit jamais constituer un objectif en soi.

Pour obtenir les résultats présentés au paragraphe suivant, certaines hypothèses de continuité et de dérivabilité des fonctions f et v doivent être posées. Cela signifie, en particulier, que les modèles à rupture doivent être considérés séparément.

3 ESTIMATION DES PARAMETRES.

Estimer les paramètres d'un modèle, c'est obtenir, à partir des observations , une certaine information sur leurs véritables valeurs. En fait, celà consiste à définir une fonction des observations $\widehat{\theta}_n = H(Y)$, appellée estimateur, qui possède de bonnes propriétés et qui apporte le maximum d'information sur les paramètres.

Les propriétés qui nous intéressent en général sont :

- L'absence de biais.

  L'espérance mathématique de l'estimateur $\widehat{\theta}_n$ est égale à θ, c'est-à-dire que si nous répétions un grand nombre de fois l'expérience avec n fixé, la moyenne des valeurs de l'estimateur obtenues tendrait vers la vraie valeur θ

  $$E_\theta \widehat{\theta}_n = \theta$$

- La consistance.

  Pour l'expérience considérée, quand le nombre d'observations n tend vers l'infini, la valeur de l'estimateur tend vers la vraie valeur de θ,

  $$\lim_{n \to +\infty} \widehat{\theta}_n = \theta$$

- L'efficacité.

  Quel que soit l'estimateur choisi, sa variance ne peut pas être plus petite qu'une certaine valeur (borne de Cramer-Rao). Si cette borne est atteinte (asymptotiquement ou non) on dit que l'estimateur est efficace.

Lorsque nous considérons les propriétés des estimateurs, une notion abondamment utilisée est celle de "résultats asymptotiques". En effet, généralement on ne peut démontrer telle ou telle propriété qu'asymptotiquement, c'est-à-dire quand $n \to +\infty$. Or, ce n'est jamais le cas et il est intéressant de connaitre la validité des résultats asymptotiques à distance finie (c'est-à-dire pour n pas trop grand) et éventuellement de recourir à d'autres méthodes.

L'information sur les paramètres est généralement apportée au travers de quantités comme le biais, la variance (ou la précision) des estimateurs et les intervalles de confiance pour les paramètres. Ces quantités n'étant pas calculables directement, diverses méthodes d'évaluation peuvent être utilisées. On distinguera donc, dans la suite de ce rapport, le choix d'un estimateur pour les paramètres, du choix de la méthode d'évaluation de ses propriétés.

On a donc recours à des méthodes d'évaluation qui peuvent être classées dans trois catégories : méthodes fondées sur les résultats asymptotiques, sur des développements ou des transformations, et sur les techniques de rééchantillonnage ou de Monte-Carlo. Le choix de la méthode d'évaluation n'est pas aisé car il n'existe pas de méthode optimale à "tous les coups".

Méthodes asymptotiques. Sous certaines conditions de régularité, les estimateurs précédemment définis suivent asymptotiquement une loi normale

$$\sqrt{n}(\widehat{\theta}_n - \theta) \sim \mathcal{N}(0, V_\theta^{-1}) \tag{2}$$

où V_θ dépend de l'estimateur considéré et est évalué pour la vraie valeur de θ.

Ce résultat peut être utilisé pour calculer la variance de l'estimateur qui est alors estimée par $\frac{1}{n}\widehat{V}_{\widehat{\theta}}^{-1}$ évaluée au point estimé. Ce résultat permet aussi de construire des intervalles de confiance et des tests d'hypothèses asymptotiques.

Méthodes d'approximation. Les résultats asymptotiques ne fournissent pas toujours de bonnes évaluations des propriétés des estimateurs, en raison de la non-linéarité plus ou moins importante (Messéan, 1984).

Ces résultats étant obtenus à partir d'un développement au premier ordre des estimateurs, on peut améliorer l'évaluation en prenant en compte les termes d'ordre supérieur. Les termes correctifs à l'ordre 2 font intervenir les mesures de non-linéarité et leur calcul n'est pas aisé. Quelques simulations effectuées pour la variance semblent indiquer que le gain n'est pas significatif (Messéan,1984).

Une autre façon de corriger les résultats asymptotiques consiste à changer de paramétrisation. En effet, la non-linéarité est fonction de la paramétrisation choisie et, pour chacun des paramètres, nous pouvons définir une transformation optimale $\phi = g(\theta)$ qui minimise la non-linéarité (Hougaard, 1982). Le résultat de normalité asymptotique s'applique également à $\widehat{\phi} = g(\widehat{\theta})$. On utilisera donc les résultats asymptotiques sur ϕ et on en déduit les propriétés de $\widehat{\theta}$ par transformation inverse (par exemple, pour construire un intervalle de confiance).

Méthodes de rééchantillonnage ou de Monte-Carlo. Ces méthodes de simulation sont très simples dans le principe et d'un coût informatique très élevé. Il s'agit de générer, par simulation numérique, des échantillons d'observations Y^* distribuées selon la même loi que les observations réelles Y. Chaque échantillon nous donne, après ajustement du modèle, une valeur $\tilde{\theta}^*$ de θ. En générant B échantillons de n valeurs de Y^* simulées, nous obtenons ainsi un B-échantillon de valeurs $\tilde{\theta}^*$ de l'estimateur de θ. On peut alors étudier les propriétés de l'estimateur à partir de ce B-échantillon.

Par exemple, la variance de l'estimateur sera estimée par la variance empirique de ce B-échantillon $\tilde{V} = \frac{1}{B}\sum_{b=1}^{B}(\tilde{\theta}^{*b} - \bar{\theta}^*)^2$, le biais par la différence entre sa moyenne empirique $\bar{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\tilde{\theta}^{*b}$ et la valeur de θ estimée à partir de l'échantillon réel. De même, des intervalles de confiance peuvent être construits en utilisant les quantiles de la distribution de ce B-échantillon.

Les méthodes de rééchantillonnage diffèrent par le mode de génération des échantillons. La méthode du Bootstrap (Efron, 1979) ne fait aucune hypothèse sur la distribution des erreurs et se fonde sur la distribution empirique des résidus $\widehat{\varepsilon}_i = y_i - f(x_i, \widehat{\theta})$ pour générer les échantillons Y^*

$$Y_i^* = f(x_i, \widehat{\theta}) + \widehat{\varepsilon}^*$$

où $\widehat{\varepsilon}^*$ est tiré de façon uniforme parmi les $\widehat{\varepsilon}_i$ (après normalisation).

3.1 Choix de l'estimateur.

Dans ce paragraphe, nous présentons quelques estimateurs qui peuvent être considérés dans le cadre de la régression non-linéaire définie par (1). Leurs propriétés ont été étudiées par S. Huet (Huet, 1986) et sont rappelées briévement ici.

Si la loi des erreurs est supposée Gaussienne, on pourra utiliser l'estimateur du maximum de vraisemblance, $\widehat{\theta}_{MV}$, qui est consistant et efficace. Il minimise la fonction $-2L_n(\theta)$ définie par :

$$-2L_n(\theta) = \sum_{i=1}^{n} \ln(v_i(\theta)) + \sum_{i=1}^{n} \frac{(y_i - f_i(\theta))^2}{v_i(\theta)}$$

avec $f_i(\theta) = f(x_i, \theta)$ et $v_i(\theta) = \sigma^2 v(x_i, \theta)$.

Les trois autres estimateurs que nous présentons ici sont utilisables quelle que soit la loi des erreurs, même si leurs propriétés en dépendent.

L'estimateur des Moindres Carrés Ordinaires, $\widehat{\theta}_{MCO}$, minimise $S_n(\theta)$ définie par :

$$S_n(\theta) = \sum_{i=1}^{n} (y_i - f_i(\theta))^2$$

$\widehat{\theta}_{MCO}$ est l'estimateur classiquement utilisé lorsque l'on parle de la méthode des moindres carrés. Il est consistant mais non efficace sauf si $v_i(\theta)$ est constant pour tout i et égal à σ^2.

L'estimateur des Moindres Carrés Pondérés, $\widehat{\theta}_{MCP}$, minimise $SW_n(\theta)$ définie par

$$SW_n(\theta) = \sum_{i=1}^{n} \frac{(y_i - f_i(\theta))^2}{v_i(\theta)}$$

$\widehat{\theta}_{MCP}$ n'est pas consistant, sauf si $v_i(\theta)$ ne dépend pas de θ, auquel cas il est également efficace.

L'estimateur des Moindres Carrés Modifiés, $\widehat{\theta}_{MCM}$, n'est pas défini par la minimisation d'un critère mais par la résolution d'un système de p équations non-linéaires

$$K_{a,n}(\theta) = \sum_{i=1}^{n} \frac{(\partial f_i(\theta)/\partial \theta^a)}{v_i(\theta)} (y_i - f_i(\theta)) = 0 \qquad a = 1, ..., p$$

$\widehat{\theta}_{MCM}$ est consistant et "plus efficace" que $\widehat{\theta}_{MCO}$, c'est-à-dire que sa matrice de variance-covariance est plus petite au sens des matrices définies positives.

Si la variance des erreurs ne dépend pas de θ, les méthodes du maximum de vraisemblance, des Moindres Carrés Pondérés ou Moindres Carrés Modifiés sont équivalentes. A priori, on n'utilisera pas la méthode des Moindres Carrés Ordinaires sauf si la variance est constante ou si les autres méthodes posent des problèmes numériques. Si on peut supposer que la loi des erreurs est gaussienne, il est préférable de choisir l'estimateur du Maximum de vraisemblance, puisqu'il est efficace. En revanche, l'estimateur des moindres carrés modifiés est plus robuste vis-à-vis de la loi des erreurs.

3.2 Evaluation des propriétés.

En général, il est impossible de calculer analytiquement la variance ou le biais d'un estimateur car l'estimateur est défini comme solution d'un problème d'optimisation non-linéaire (minimisation d'une distance non quadratique ou résolution d'un système d'équations non-linéaires) et qu'il ne s'exprime donc pas sous forme analytique.

Ce procédé de simulation confère à la méthode du Bootstrap une certaine robustesse vis-à-vis de la distribution des erreurs mais pose des problèmes de biais lorsque le nombre d'observations est faible. Si la loi des erreurs est parfaitement connue, la génération se fera de préférence selon cette loi.

Il existe de multiples variantes de ces méthodes qui connaissent, aujourd'hui, un incontestable succès. Une étude spécifique à la régression non-linéaire souligne, néanmoins, les nombreuses difficultés que soulève l'application de ces méthodes dans le contexte de la régression non-linéaire (Huet S., Jolivet E, Messéan A. & Nicole O., 1986).

3.3 Aspects numériques.

La résolution numérique du problème d'estimation se ramène à la résolution d'un système d'équations non-linéaires qui correspond aux équations normales, c'est-à-dire les dérivées par rapport aux paramètres du critère à minimiser. Ces équations normales ont toujours la forme générale suivante

$$B^T(\theta)(Z_Y - \eta(\theta)) = 0 \qquad (3)$$

où Z_Y est un vecteur fonction des observations (les statistiques exhaustives dans certains cas), $\eta(\theta)$ est l'espérance de Z_Y, ou la limite de Z_Y et $B(\theta)$ est une matrice $h \times p$ où h est la dimension de Z_Y.

Par exemple, l'estimateur classique des Moindres Carrés est défini par

$$\begin{aligned} Z_i &= y_i \\ \eta_i(\theta) &= f(x_i, \theta) \qquad i = 1, ..., n \\ B_{i,a}(\theta) &= \frac{\partial f_i(\theta)}{\partial \theta^a} \end{aligned}$$

Chaque estimateur est ainsi défini par les expressions de B, Z et η. La résolution du système (3) peut donc se faire à l'aide des algorithmes de Gauss-Newton ou de Gauss-Marquardt dont le principe est le suivant : on linéarise $\eta(\theta)$ autour d'une valeur initiale θ_0 et on obtient le système linéaire

$$B^T(\theta_0)(Z_Y - \eta(\theta_0)) = B^T(\theta_0)D(\theta_0)(\theta - \theta_0)$$

où $D(\theta_0)$ est la matrice des dérivées de η par rapport à θ.

L'inversion directe de ce système nous donne la nouvelle approximation de Gauss-Newton. L'approximation de Gauss-Marquardt, quant à elle, est obtenue en ajoutant à la diagonale de $B^T(\theta_0)D(\theta_0)$ un certain scalaire λ. Ce processus est répété jusqu'à convergence vers la solution (Messéan A., 1986).

La prise en compte de contraintes du type $\theta_{inf} < \theta < \theta_{sup}$ ou $\theta^a = \theta^b$ est aisément effectuée à l'aide d'une reparamétrisation (Messéan, 1986).

Le bon comportement des algorithmes suppose que la fonction à minimiser ne soit pas trop irrégulière et que nous disposions de valeurs intiales convenables. En général, la signification biologique des paramètres apporte une information non négligeable (signe des paramètres, intervalle de variation, ...) et facilite la convergence du processus numérique.

4 VALIDATION DU MODELE.

La validation du modèle consiste à mettre en oeuvre un certain nombre de techniques destinées, d'une part, à vérifier le bien-fondé des hypothèses posées (formes de l'espérance et de la variance, loi des erreurs) et, d'autre part, à apprécier l'adéquation du modèle considéré par rapport aux objectifs fixés.

Analyse des résidus. Ce point essentiel s'effectue principalement à l'aide de graphiques et de tests statistiques :

- Les graphiques permettent de détecter visuellement des défauts d'adéquation du modèle par la mise en évidence de tendances, et également d'étudier la variabilité de la variance.
- Les tests (paramétriques ou non-paramétriques) permettent de préciser l'analyse subjective donnée par les graphiques et sont importants pour tester l'indépendance des erreurs ou leur normalité.

Cette analyse ne diffère pas de celle rencontrée dans le cadre de la régression linéaire et une bibliographie abondante existe (Cook and Tsai, 1985; Draper and Smith, 1981).

Tests d'hypothèses. Si l'on dispose de répétitions, c'est-à-dire si la variable Y est observée plusieurs fois pour une même combinaison des variables explicatives, on pourra effectuer un test d'adéquation du modèle aux données dont le principe est de comparer l'écart du modèle aux données à la variabilité de ces données. Il s'agit, en fait, d'un test de sous-modèle au même titre que les tests classiques d'hypothèses sur les paramètres.

Pour l'ensemble des tests d'hypothèses emboîtées et si le modèle statistique peut être considéré comme gaussien, le test de rapport de vraisemblance constitue une méthode simple dans le principe et robuste vis-à-vis de la non-linéarité : en effet, il est invariant par changement de paramétrisation (Messéan, 1982).

Son principe est le suivant. On veut comparer un modèle M_1 à q paramètres au modèle M_2 à p paramètres ($q < p$, c'est-à-dire M_1 emboîté dans M_2), M_2 étant supposé être un modèle s'ajustant correctement aux données. Soit $L_1(Y)$ le maximum de vraisemblance obtenu pour M_1, et $L_2(Y)$ celui obtenu pour M_2. Alors $-2\ln(L_1(Y) - L_2(Y))$ est asymptotiquement distribué selon une loi de χ^2 à $p-q$ degrés de liberté.

Pratiquement, il suffit de décrire ses sous-hypothèses sous forme de contraintes sur les paramètres, d'ajuster à nouveau le modèle sous ces contraintes et de comparer la différence des vraisemblances obtenues à un $\chi^2_\alpha(p-q)$ pour un test au niveau α (Huet & Messéan, 1986).

Analyse de sensibilité. Une autre technique permet de juger de l'influence du plan d'expérience (le choix des valeurs x_i où sont effectuées les mesures). Elle consiste à représenter graphiquement les fonctions de sensibilité, c'est-à-dire les dérivées de $f(x_i, \theta)$ par rapport à chacun des paramètres θ, normalisées par l'écart type de l'observation

$$S_a(x,\theta) = \frac{\partial f(x,\theta)/\partial\theta^a}{v(x,\theta)} \qquad a = 1,...,p$$

Un paramètre est d'autant plus "favorisé" (en termes de précision) que le plan d'expérience contient de points x_i pour lesquels la valeur absolue de $S_a(\theta)$ est grande. Celà permet de conseiller éventuellement un nouveau plan d'expérience qui privilégie tel ou tel paramètre en fonction de l'objectif de l'étude.

5 INTERVALLES DE CONFIANCE.

Les méthodes de construction d'intervalles de confiance sont, en fait, directement dérivées des méthodes d'évaluation des propriétés des estimateurs abordées en 3.2.

On peut, tout d'abord, exploiter les résultats asymptotiques (2) en utilisant l'écart-type "asymptotique" $\widehat{\sigma}_{\widehat{\theta}}$ associé à chaque paramètre. Un intervalle au niveau α est alors donné par

$$[\widehat{\theta} - \widehat{\sigma}_\theta \; t_{n-p;1-\alpha}, \widehat{\theta} + \widehat{\sigma}_\theta \; t_{n-p;1-\alpha}]$$

où $t_{n-p;1-\alpha}$ est la valeur critique au niveau $1-\alpha$ d'une loi de Student à $n-p$ degrés de liberté.

Des intervalles de confiance symétriques et de même nature peuvent être construits en utilisant n'importe quelle autre estimation de l'écart type de $\widehat{\theta}$.

Les méthodes de rééchantillonnage permettent de construire deux types d'intervalles de confiance : le premier consiste à utiliser des estimations de la variance de l'estimateur pour bâtir un intervalle symétrique comme pour le résultat asymptotique, le second utilise les quantiles de la distribution de l'estimateur obtenue par simulation.

L'échantillon de B valeurs de $\widehat{\theta}^*$ obtenu par simulation peut être représenté sur un axe et, en éliminant un pourcentage $\frac{\alpha}{2}$ de valeurs aux extrémités droite et gauche, on obtient les bornes d'un intervalle de confiance au niveau $1-\alpha$. Il existe de nombreuses variantes de cette méthode qui permettent d'obtenir un meilleur intervalle (Efron, 1985).

Si la vraisemblance est gaussienne, le test de rapport de rapport de vraisemblance permet de construire des intervalles de confiance qui réalisent un excellent compromis (Messéan, 1984) entre la facilité de construction et la probabilité de recouvrement (probabilité que la valeur exacte du paramètre appartienne à l'intervalle de confiance).

Enfin, une méthode itérative simple à mettre en oeuvre utilise la transformation optimale $\phi = g(\theta)$ réduisant la non-linéarité pour construire un intervalle de confiance (Hougaard, 1982).

Une présentation détaillée de ces différentes méthodes accompagnée d'une comparaison à partir de simulations numériques est donnée dans Huet et al., 1986.

6 EXEMPLE.

Dans cette section, nous illustrons certains des aspects statistiques précédemment évoqués par l'analyse de données radioimmunologiques à l'aide du logiciel CS-NL, développé au laboratoire de Biométrie de l'INRA-Jouy (Bouvier et al., 1985; Huet, S. & Messéan, A., 1986). Les données sont des réponses radioactives à des doses croissantes de Cortisol et sont reportées dans la table 1.

Le modèle utilisé est celui de Richards avec une variance proportionnelle au carré de l'espérance

$$\begin{aligned} f(x,\theta) &= n + \frac{d-n}{(1+\exp(a+bx))^g} \quad \text{avec } \theta^T = (n,d,a,b,g) \\ v(x,\theta) &= \sigma^2 f^2(x,\theta) \end{aligned}$$

La forme de la variance retenue est suggérée par l'étude des variations des variances empiriques en fonction des moyennes qui est représentée sur la figure 1.

L'estimation des paramètres par la méthode du maximum de vraisemblance (en supposant la distribution normale) nous donne les résultats de la Table 2 (sortie du programme CS-NL).

La figure 2 représente le graphe des observés et des ajustés par rapport à la variable explicative et la table 3 donne une représentation simple des résidus normalisés.

Ces graphiques montrent que l'adéquation du modèle aux données est visuellement satisfaisante. Cette impression peut être vérifiée à l'aide d'un test d'adéquation du modèle puisque nous disposons de répétitions. On suppose que le modèle de Richards est un sous-modèle du

Log-concentration	Réponses (cpm)
-2.000	2868 2785 2849 2805 2779 2588 2701 2752
-1.699	2615 2651 2506 2498
-1.398	2474 2573 2378 2494
-1.222	2152 2307 2101 2216
-1.097	2114 2052 2016 2030
-1.000	1862 1935 1800 1871
-0.699	1364 1412 1377 1304
-0.398	910 919 855 875
-0.222	702 701 689 696
-0.097	586 596 561 562
0.000	501 495 478 493
0.176	392 358 399 394
0.301	330 351 343 333
0.602	250 261 244 242
1.000	131 135 134 133

Table 1: Dosage radioimmunologique du Cortisol : réponses observées.

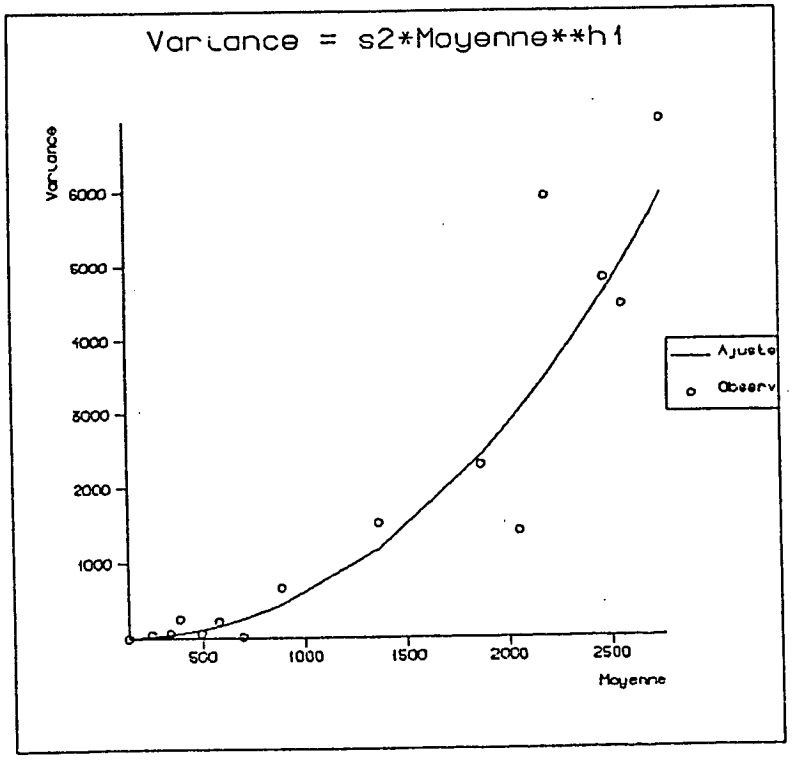


Figure 1: Variabilité des observations par rapport à leurs moyennes.

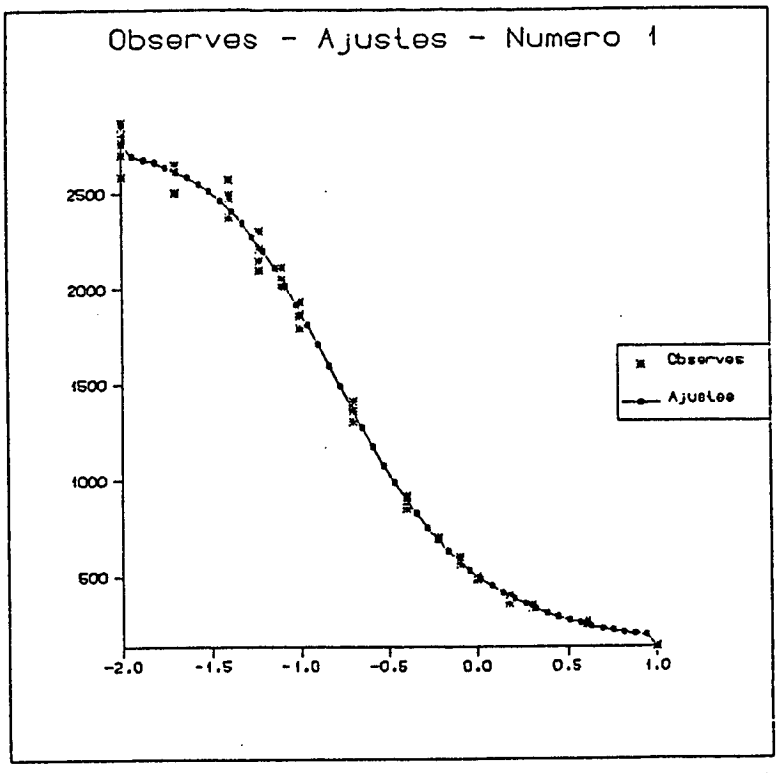


Figure 2: Dessin des ajustés et observés.

modèle général $y_i = \mu_i + \varepsilon_{ij}$, où les paramètres sont les μ_i (au nombre de 15, le nombre de combinaisons de variables explicatives). L'ajustement de ce modèle (toujours en supposant la variance proportionnelle au carré de l'espérance, c'est-à-dire $v_i = \sigma^2 \mu_i^2$), nous donne une vraisemblance égale à 596.6 avec 49 degrés de liberté (ddl). L'ajustement du modèle de Richards ayant une vraisemblance de 607.9 avec 59 ddl, la différence de 11.3 est à comparer à la valeur critique d'un χ^2 à 59-49=10 ddl. Cette différence n'est pas significative et le modèle de Richards peut être considéré comme s'ajustant correctement aux données.

Une autre question est de savoir si le modèle logistique suffit à représenter les données. Le modèle logistique étant un cas particulier du modèle de Richards avec $g = 1$ (pas de dissymétrie), il suffit de tester l'hypothèse $g = 1$. Pour celà, nous ajustons à nouveau le modèle sous la contrainte $g = 1$. La vraisemblance obtenue est 647.6 avec 60 ddl et la différence par rapport au modèle complet est donc de 647.6 - 607.9 = 39.7 à comparer à un χ^2 à 1 ddl. Cette hypothèse est donc nettement rejetée et le modèle ne peut être considéré comme symétrique.

Enfin, un intervalle de confiance peut être construit pour chacun des paramètres en utilisant encore une fois le test de rapport de vraisemblance (Huet, S. & Messéan, A., 1986). Ces intervalles sont reportés dans la table 4 et sont meilleurs que ceux qui dérivent directement de l'écart-type associé à chaque paramètre (voir (2)).

	no_param	estime	ecart_type	p1	p2	p3	p4	p5
p1	1.000	133.423	1.949	1.000	0.027	-0.177	-0.132	0.214
p2	2.000	2758.695	26.317	0.027	1.000	-0.430	-0.567	0.488
p3	3.000	3.201	0.223	-0.177	-0.430	1.000	0.975	-0.995
p4	4.000	3.262	0.160	-0.132	-0.567	0.975	1.000	-0.978
p5	5.000	0.608	0.041	0.214	0.488	-0.995	-0.978	1.000

estimateur : maximum de vraisemblance
mod_var : yes
sigma2 : resid = 8.68945E-04
-2*Log(V) = 6.07928E+02 avec 59 ddl

Table 2: Argument de retour du programme d'estimation de CS-NL.

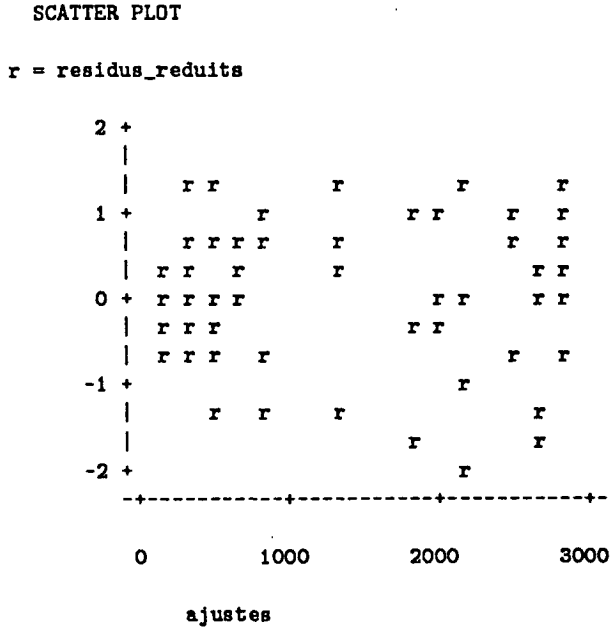


Table 3: Graphe des résidus réduits en fonction des ajustés.

```
intervalles_conservatifs

              borne_sup.
    borne_inf.

P1     126.970   140.533
P2    2671.158  2853.252
P3       2.511     4.071
P4       2.782     3.904
P5       0.479     0.766

estimateur : maximum de vraisemblance
mod_var : yes
sigma2 : resid =  8.68945E-04
-2*Log(V) =  6.07928E+02 avec    59 ddl
```

Table 4: Intervalles de confiance pour les paramètres.

REFERENCES.

BOUVIER A. and al. (1985). *CS-NL : manuel d'utilisation.* Technical Report, Laboratoire de biométrie, INRA.

COOK, D. and TSAI, C. (1985). "Residuals in nonlinear regression" *Biometrika*, 72, 1, p. 23-29.

DRAPER, N. and SMITH, H. (1981). *"Applied regression analysis"* Wiley Eds, New-York, 710p.

EFRON, B. (1979). "Bootstrap methods : another look to Jackknife" *Annals of Statistics*, 7, p. 1-26.

EFRON, B. (1985). *"Better Bootstrap confidence intervals"* Technical Report. Dept. Stat. Stanford University.

HOUGAARD, P. (1982). "Parametrization of nonlinear models" *J. R. Statist. Soc. B*, 44(2), p. 244-252.

HUET, S. (1984). "Dosages radioimmunologiques : quelle analyse statistique?" *Recherche, Nutrition, Développement*, 24(3), p. 209-219.

HUET, S., JOLIVET, E., MESSÉAN, A. et NICOLE, O. (1986). *"Some simulation results about confidence intervals construction and Bootstrap methods"* Technical Report. INRA, Laboratoire de Biométrie du CRJJ.

HUET S. et MESSÉAN A. (1986) NL : A statistical package for general nonlinear regression problems. IN *Proceedings in computational statistics*, (ed. F. DE ANTONI, N. LAURO and A. RIZZI), Physica Verlag, Heidelberg Wien, 326-331

HUET S. (1986) Maximum Likelihood and Least Squares Estimators for a Non-Linear Model with Heterogeneous Variances. *Statistics*, **17**, 517-526.

JOLIVET, E. (1982). *"Modèles mathématiques en Biologie"* Actualités scientifiques et agronomiques, Masson, 151p.

LEBRETON, J.-D. et MILLIER, C. (1982). *"Modèles dynamiques déterministes en Biologie"* Masson, 208p.

MESSÉAN, A. (1982). "Régions de confiance dans le modèle non-linéaire" *Journal de la Société de Statistique de Paris*, 123, 2, p. 134-143

MESSÉAN A. (1984). *Application de la géométrie différentielle à la statistique du modèle non-linéaire.* Thèse de Doctorat-ingénieur, Université Paris-Sud (Orsay), 70p.

MESSÉAN A., NICOLE O. et VILA J.P. (1984). HAUSS 82: a New Tool for Nonlinear Fitting and the Study of Nonlinear Models. IN *Proceedings in computational statistics*, (ed. T. HAVRANEK, Z. SIDAK and M. NOVAK), Physica Verlag, Heidelberg Wien, 434-439.

MESSÉAN A. (1986) A generalization of Gauss-Newton and Gauss-Marquardt algorithms for estimation in exponential families. *Soumis pour publication à CSQ.*

Redressabilité des champs quadratiques plans sans singularités

Laurent BARATCHART
Eric BENOIT
José GRIMM

INRIA Sophia-Antipolis
Avenue Emile Hugues
06560 Valbonne

1. Introduction

Les systèmes quadratiques plans ont fait l'objet d'études intenses depuis des dizaines d'années (cf [1] et sa bibliographie) et le 16e problème de Hilbert concerne en partie le nombre maximum de cycles limites de telles équations. Les champs quadratiques sans point singulier (et donc sans cycle limite) sont naturellement peu intéressants de ce point de vue, et leur comportement dynamique assez trivial en général a suscité relativement peu d'intérêt. Il est en effet connu (cf par exemple [2,3]) que ces champs sont redressables en général, la seule obstruction étant l'existence "accidentelle" d'une "composante de Reeb" dans le portrait de phase. Le modeste objectif du présent travail est de fournir une preuve élémentaire (i.e. n'utilisant que le théorème de Poincaré-Bendixon et un peu de géométrie) qu'un tel champ est génériquement redressable de façon particulière, et d'étudier tous les cas non génériques. Dès lors cet article s'apparente beaucoup à une classification, à un changement de coordonnées linéaire près, des systèmes quadratiques plans sans singularités. Une telle classification existe pour les champs homogènes, cf [4,5]. On en a déduit une procédure informatique décidant de la redressabilité d'un tel champ, qui nous a servi pour faire les illustrations.

2. Préliminaires et notations

Définition 1. *Etant donné un champ de vecteurs $\Phi : \mathbf{R}^2 \to \mathbf{R}^2$, on dit que Φ est redressable (par Ψ) s'il existe un difféomorphisme Ψ de $\mathbf{R}^2$ dans un ouvert U de $\mathbf{R}^2$ tel que $D\Psi.\Phi = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$.*

Il est alors clair que si on fait le changement de variables $(u, v) = \Psi(x, y)$ dans le système différentiel $\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \Phi(x, y)$ on obtient le système $\dot{u} = 1$, $\dot{v} = 0$.

Définition 2. *On dit qu'un champ de vecteurs $\Phi : \mathbf{R}^2 \to \mathbf{R}^2$ est redressable par une droite s'il existe une droite qui coupe une fois et une seule et de façon transverse toute trajectoire du flot associé.*

Il est facile de voir que *redressable par une droite* entraîne *redressable*. Par changement de variables on peut toujours supposer que la droite est l'axe des y. Définissons une fonction Ψ de la façon suivante. Si (x, y)

est un point du plan, γ la trajectoire du champ ayant comme condition initiale (x, y) pour $t = 0$, v l'ordonnée du point d'intersection de la droite et de γ, u le temps auquel γ passe par ce point, alors $\Psi(x, y) = (-u, v)$.

Soit $\varphi_t(x, y)$ le flot, i.e. $\varphi_t(x, y)$ est obtenu en intégrant le champ pendant le temps t à partir de la condition initiale (x, y), donc

$$\frac{d}{dt}\varphi_t(x, y) = \Phi(\varphi_t(x, y))$$

Il est connu (cf [6]) que φ_t est un difféomorphisme local et on voit facilement que sa dérivée $D\varphi_t$ est solution de l'équation différentielle

$$\frac{d}{dt}D\varphi_t(x, y) = D\Phi(\varphi_t(x, y)) \circ D\varphi_t(x, y).$$

Il s'ensuit aisément que $D\varphi_t(x, y).\Phi(x, y) = \Phi(\varphi_t(x, y))$ car ils vérifient la même équation différentielle avec les mêmes conditions initiales.

Soit $\phi(u, v) = \varphi_u(0, v)$. Alors , en posant $(x, y) = \varphi_u(0, v)$

$$\begin{cases} \dfrac{\partial \phi}{\partial u} = \Phi(x, y) = D\varphi_u(0, v).\Phi(0, v) \\ \dfrac{\partial \phi}{\partial v} = D\varphi_u(0, v).\begin{pmatrix} 0 \\ 1 \end{pmatrix}. \end{cases}$$

Comme $D\varphi_u$ est inversible, et que $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ et $\Phi(0, v)$ sont indépendants car le champ est transverse le long de la droite, on voit que ϕ est un difféomorphisme local. Comme on a de façon évidente $\phi \circ \Psi = I$, on déduit $D\Psi = D\phi^{-1}$, donc, comme $D\phi.\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \dfrac{\partial \phi}{\partial u} = \Phi$ on a bien $D\psi.\Phi = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$.

Notons que si toute trajectoire coupe la droite de façon transverse, elle ne peut la couper plus d'une fois, car sur la droite le champ est dirigé vers l'un des demi-plans.

Notons aussi que s'il existe une courbe Γ telle que toute trajectoire coupe Γ transversalement, et qu'il existe un difféomorphisme de $\mathbf{R}^2$ dans $\mathbf{R}^2$ transformant la courbe Γ en une droite, le champ se redresse (il suffit de faire ce changement de variables).

En particulier, si Γ est une branche d'hyperbole, il existe un difféomorphisme de $\mathbf{R}^2$ dans $\mathbf{R}^2$ transformant Γ en une droite, car on peut toujours écrire, modulo un changement de variables la branche d'hyperbole sous la forme $y = f(x)$ et dans ce cas le difféomorphisme $(x, y) \to (x, y - f(x))$ convient.

On examinera dans la suite les champs quadratiques plans sans singularités c'est-à-dire ceux qui s'écrivent

$$\begin{cases} \dot{x} = ax^2 + bxy + cy^2 + dx + ey + f \\ \dot{y} = a'x^2 + b'xy + c'y^2 + d'x + e'y + f' \end{cases}$$

où $\dot{x}$ et $\dot{y}$ ne peuvent s'annuler en même temps.

On dira qu'une propriété concernant de tels champs est générique si elle est vraie pour toutes les valeurs des coefficients a, b, c etc., qui ne sont pas racines d'un certain polynôme non nul.

Lorsqu'on fait un changement de notations, il sera de la forme :

$$\begin{cases} X = \alpha x + \beta y + \gamma \\ Y = \alpha' x + \beta' y + \gamma' \\ T = \nu t. \end{cases}$$

Si l'on ne précise pas la valeur de X, par exemple, c'est qu'il s'agit de $X = x$.

Le nouveau champ sera noté :

$$\begin{cases} \dot{X} = AX^2 + BXY + CY^2 + DX + EY + F \\ \dot{Y} = A'X^2 + B'XY + C'Y^2 + D'X + E'Y + F'. \end{cases}$$

Lors de chaque changements de variables effectués en cascade, il nous arriivera cependant de reprendre les notations x et y sitôt le changement de variables effectué. Lorsqu'on considère une trajectoire du champ, il s'agit toujours d'une solution maximale de l'équation différentielle, qu'on suppose définie sur l'intervalle $]T_-, T_+[$.

L'objet de cet article est de montrer le théorème suivant :

Théorème. *Génériquement, un champ quadratique plan sans singularités se redresse par une droite. De façon précise, les seuls champs qui se redressent mais pas par une droite sont ceux qui peuvent se ramener par une transformation linéaire aux équations (CP3), (CP4), (CP5), (CP2b) avec $a \geq 0$, (CP2c) avec $ab \geq 0$ et $a \neq b$, (CP0) avec $0 < a \leq 1$. Les champs qui ne se redressent pas sont ceux qui se ramènent aux équations (CP2a), (CP2b) avec $a < 0$, (CP2c) avec $ab < 0$, (CP0) avec $1 < a < 0$.*

3. Résultats préparatoires

Lemme 1. *Il n'existe pas de compact positivement invariant par le flot d'un champ plan sans singularités.*

Ceci est essentiellement le théorème de Poincaré-Bendixon [7].

Lemme 2. *Considérons un champ quadratique plan sans singularités et une droite. Une trajectoire ne peut couper la droite en plus de trois points. Par conséquent, si, sur une trajectoire, x n'est pas borné, c'est que x tend vers $+\infty$ ou x tend vers $-\infty$. De même, si x ne tend pas vers 0, 0 ne peut être valeur d'adhérence de $\{x(t) |\ t \in]T_-, T_+[\}$.*

Preuve. On peut toujours se ramener au cas où la droite est $x = 0$. Remarquons tout d'abord que les points d'intersection d'une trajectoire avec cette droite sont tous distincts car il ne peut y avoir de cycle limite sans point fixe à l'intérieur. Observons ensuite que si la trajectoire traverse deux fois l'axe dans le même sens en des points entre lesquels $\dot{x}$ ne change pas de signe, cela constitue un piège (i.e. un compact invariant par le flot) en temps positif ou négatif suivant les cas. A présent le signe de $\dot{x}$ est gouverné par un trinôme en y le long de la droite.

- Si $\dot{x}$ ne s'annule pas, la trajectoire ne peut couper la droite plus d'une fois.

- Si $\dot{x}$ a une racine double, $\dot{x}$ ne change pas de signe, et la trajectoire ne peut couper plus de deux fois, car c'est seulement aux points où $\dot{x}$ s'annule que l'on peut traverser dans le sens opposé à celui du champ.

- Si $\dot{x}$ n'a que des racines simples, et si la trajectoire renncontre la droite $x = 0$ en un point où $\dot{x} = 0$, il découle immédiatement du fait que les racines sont simples, que $\ddot{x} \neq 0$, de sorte que la trajectoire traverse l'axe. Elle traverse donc l'axe à chaque fois qu'elle le rencontre, et si elle le rencontre plus de trois fois, on a nécessairement deux traversées dans le même sens dans une zone où $\dot{x}$ est de signe constant, ert donc un piège.

Les autres affirmations découlent de ce qui précède.

Lemme 3. *Considérons une trajectoire d'un champ quadratique plan. Supposons $c \neq 0$ ou $c' = 0$. Si $\dot{x} \to +\infty$ (resp. $-\infty$) lorsque $t \to T_+$ on aura $x \to +\infty$ (resp $-\infty$) lorsque $t \to T_+$.*

Preuve. Il suffit de considérer le cas $\dot{x} \to +\infty$. On suppose la solution maximale définie sur $]T_-, T_+[$. Il existe alors t_1 tel que si $t > t_1$ on ait $\dot{x} > 1$. Si la conclusion du lemme était fausse, comme x est croissant pour $t > t_1$, x serait borné et donc T_+ serait fini. De plus (lemme 2, où x est remplacé par y) $|y| \to +\infty$. Il existerait donc t_2 tel que si $t > t_2$ on ait $y > 1$ ou $y < -1$.

Etudions d'abord le cas $c' = 0$. Alors $\dot{y} = (b'x + e')y + a'x^2 + d'x + f'$. Si $y \to -\infty$, on remplace y par $-y$, ce qui ne va pas changer la forme de $\dot{y}$. Comme $T_+ < \infty$ et $y \to +\infty$ on a $\dot{y} \to +\infty$ et donc $b'x + e' > 0$ pour $t > t_3$. On suppose $t > t_3$. Comme x est borné il existe A et $B > 0$ tels que si $t_3 < t < T_+$, $a'x^2 + d'x + f' < A$ et $b'x + e' < B$. Alors $\dot{y} < By + A$, d'où $y < Ke^{Bt} - A/B$ pour une certaine constante K, (car si $w = y - Ke^{Bt} + A/B$ on a $\dot{w} < Bw$, et il suffit de choisir $K = y_0 + A/B$) et ceci contredit le fait que y est non borné.

Supposons maintenant $c' \neq 0$. Alors $c \neq 0$. Dans ce cas si $t \to T_+$, $\dot{x} \sim cy^2$ et $\dot{y} \sim c'y^2$. Il existe donc t_4 et $K > 0$ tel que si $t > t_4$ on ait $\dot{x} > K|\dot{y}|$, et $\dot{y}$ est de signe constant. Comme y est non borné, on en déduit que x est non borné. CQFD.

Cas général

$$\begin{cases} \dot{x} = x^2 - y^2 + 2x + 3y - 4 \\ \dot{y} = x^2 + y^2 + 1 \end{cases}$$

Théorème 1. *Considérons un champ quadratique plan sans singularités, vérifiant* $b^2 - 4ac < 0$. *Alors le champ se redresse par une droite.*

Preuve : Comme $b^2 - 4ac < 0$, on a $c \neq 0$ et le lemme 3 s'applique. De plus la courbe $\dot{x} = 0$ est une ellipse (éventuellement vide), donc est compacte. Il existe donc un carré $|x| < k$, $|y| < k$ en dehors duquel $\dot{x}$ est de signe constant, que l'on peut supposer positif, sans rien changer à la suite. Alors si $t \to T_{\pm}$ on a $x \to \pm\infty$, car si x était borné, on aurait $y \to \infty$ d'où $\dot{x} \to +\infty$, ce qui contredit le lemme 3. En particulier toute trajectoire coupe toute droite $x = C$, et si C est bien choisi (par exemple $|C| > k$) l'intersection est transverse. CQFD.

Notons qu'une valeur convenable pour k est par exemple :

$$1 + \left|\frac{2bd - 4ae}{b^2 - 4ac}\right| + \left|\frac{d^2 - 4af}{b^2 - 4ac}\right|.$$

Théorème 2. *Considérons un champ quadratique plan sans singularités, vérifiant il existe* $k > 0$ *tel que* $|\dot{x}| > k$ *et* $c \neq 0$ *ou* $c' = 0$. *Alors le champ se redresse par une droite.*

Preuve. Le champ se redresse par les droites $x = C$ où C est quelconque. En effet, la droite est transverse par hypothèse à toutes les trajectoires du champ. Il reste à voir que toute trajectoire la coupe. Pour cela il suffit simplement de montrer que si $t \to T_+$ on a $x \to \pm\infty$, où le signe $\pm$ est le signe de $\dot{x}$, car on en déduit alors en inversant le temps que x parcourt $]-\infty, +\infty[$ le long de chaque trajectoire. D'après le lemme 3, ceci est vrai si $\dot{x} \to \pm\infty$. Ceci est également manifeste si T_+ est infini.

Supposons x borné et T_+ fini. Alors, si c est non nul, et comme $|y| \to +\infty$, $|\dot{x}| \to +\infty$, et le lemme 3 donne une conclusion. Si $c = 0$, alors $c' = 0$ par hypothèse et la preuve du lemme 3 montre que l'hypothèse T_+ fini est absurde. CQFD.

Faisons une première transformation du champ. On pose $Y = y - \alpha x$. Alors $A' = a' + (b' - a)\alpha + (c' - b)\alpha^2 - c\alpha^3$.

Si $c \neq 0$, alors $A' = 0$ a toujours une solution en α.

Si $c = 0$, en posant $X = y$, $Y = x$, on obtient de même $A' = 0$.

On pose $\Lambda = A(AC'^2 + B'^2C - BB'C')$. Notons que Λ est le résultant

$$(AC' - A'C)^2 - (AB' - A'B)(BC' - B'C)$$

de la partie homogène des équations $\dot{X}$ et $\dot{Y}$. Toute transformation linéaire inversible multiplie Λ par un nombre strictement positif. Nous allons étudier dans la suite le cas $AB'\Lambda \neq 0$. Ce cas est générique, car la quantité en question est un polynôme en les variables a, b, c, a', b', c' et α, où α est racine d'un autre polynôme. Eliminant α entre ces deux polynômes, on obtient un polynôme P, dépendant de a, b, c, a', b' et c', tel que si P et c sont non nuls, on est dans le cas $AB'\Lambda \neq 0$, quelle que soit la racine α choisie. Ceci justifie le titre de la section suivante.

On va donc étudier dans la suite les systèmes de la forme :

$$\begin{cases} \dot{x} = ax^2 + bxy + cy^2 + dx + ey + f \\ \dot{y} = b'xy + c'y^2 + d'x + e'y + f'. \end{cases}$$

4. Etude du cas générique $\Lambda ab' \neq 0$

4.1. Existence de point fixe si $\Lambda < 0$

On pose $X = x + \alpha$, $Y = y + \beta$, avec $\alpha = \frac{e'b' - 2c'd'}{b'^2}$ et $\beta = \frac{d'}{b'}$.

Dans ce cas on obtient:

$$\begin{cases} \dot{X} = aX^2 + bXY + cY^2 + DX + EY + F \\ \dot{Y} = b'XY + c'Y^2 + F'. \end{cases}$$

On peut donc supposer $d' = e' = 0$, sans changer la valeur de Λ.

Ecrivons les équations aux points fixes $\dot{x} = \dot{y} = 0$, et multiplions la première relation par y^2. On obtient:

$$\begin{cases} a(xy)^2 + by^2(xy) + cy^4 + Dy(xy) + Ey^3 + Fy^2 = 0 \\ xy = -\dfrac{c'y^2 + F'}{b'}. \end{cases}$$

Reportant la deuxième équation dans la première on obtient, après multiplication par ab'^2,

$$\Lambda y^4 + A_1 y^3 + A_2 y^2 + A_3 y + a^2 F'^2 = 0$$

où A_1, A_2, A_3 sont certaines quantités sans importance.

Si $F' \neq 0$ cette équation est l'équation aux ordonnées des points fixes, puisque $y = 0$ n'est pas racine; comme $\Lambda < 0$, elle a une racine au moins.

Si $F' = 0$ on a

$$\begin{cases} \dot{x} = ax^2 + bxy + cy^2 + Dx + Ey + F \\ \dot{y} = (b'x + c'y)y. \end{cases}$$

Les équations aux points fixes sont donc :

$$\text{soit} \qquad y = 0, \quad ax^2 + Dx + F = 0$$

$$\text{soit} \qquad x = -\frac{c'}{b'}y, \quad \Lambda y^2 + A_4 y + b'^2 aF = 0.$$

Si $aF \leq 0$ la première équation a une solution. Si $aF > 0$ c'est la deuxième qui a des solutions.

Donc, si $\Lambda < 0$, il y a un point fixe.

4.2. Redressabilité du cas $\Lambda > 0$

Posons $u = x + \alpha y$. Alors $\dot{u} = Au^2 + Buy + Cy^2 +$ des termes linéaires avec

$$A = a \quad B = b + \alpha(b' - 2a) \quad C = c + \alpha(c' - b) + \alpha^2(a - b')$$

et

$$\Delta = B^2 - 4AC = b'^2\alpha^2 + (2bb' - 4ac')\alpha + b^2 - 4ac.$$

Le discriminant de Δ en α est 16Λ.

Comme $\Lambda > 0$ on peut choisir α de sorte que $\Delta < 0$, et donc le théorème 1 s'applique : le champ se redresse.

Ceci prouve d'ores et déjà la partie générique du théorème principal. La suite de ce travail consiste à se ramener à un certain nombre de cas particuliers qui seront étudiés un à un.

5. Etude des cas non génériques

5.1. Etude du cas $\Lambda = 0$, $ab' \neq 0$

Posons $X = b'x + c'y$. Alors

$$\begin{cases} \dot{X} = AX^2 + BXY + DX + EY + F \\ \dot{Y} = XY + D'X + E'Y + F' \end{cases}$$

avec $A = a/b' \neq 0$. On pose ensuite $X' = X + E'$ et $Y' = Y + D'$ pour annuler E' et D'.

On est donc ramené à l'étude d'un système du type

$$\begin{cases} \dot{x} = ax^2 + bxy + dx + ey + f \\ \dot{y} = xy + f'. \end{cases}$$

Comme il n'y a pas de points fixes, on vérifie que $e = 0$.

Si $b = f' = 0$, alors $ax^2 + dx + f$ est de signe constant (puisqu'il n'y a pas de points fixes) et le champ se redresse par le théorème 2.

Si $b = 0$, $f' \neq 0$, et si $ax^2 + dx + f$ n'a pas de racines, on conclut comme précédemment. S'il a une racine non nulle, il y a un point fixe, sinon, en posant $Y = -\dfrac{y}{f'}$ on se ramène à :

$$\begin{cases} \dot{x} = ax^2 \\ \dot{y} = xy - 1. \end{cases} \qquad (CP0)$$

Si $b \neq 0$, on pose $X = -x + by$, $Y = x$ et on obtient

$$\begin{cases} \dot{X} = -aY^2 - dY + (bf' - f) \\ \dot{Y} = XY + (a+1)Y^2 + dY + f \end{cases}$$

qui est de la forme

$$\begin{cases} \dot{x} = cy^2 + ey + f \\ \dot{y} = xy + c'y^2 + d'x + e'y + f' \end{cases} \qquad (E0)$$

avec $c \neq 0$ qui sera étudié plus loin (dans ce cas particulier $d' = 0$).

5.2. Etude du cas $a \neq 0$, $b' = 0$, $c' \neq 0$

On étudie

$$\begin{cases} \dot{x} = ax^2 + bxy + cy^2 + dx + ey + f \\ \dot{y} = c'y^2 + d'x + e'y + f'. \end{cases}$$

On pose $X = x + \alpha y$.

Alors $A = a$, $B = b - 2a\alpha$, $C = a\alpha^2 - b\alpha + c + \alpha c'$ et $\Delta = B^2 - 4AC = b^2 - 4ac - 4ac'\alpha$.

Comme $ac' \neq 0$, on peut choisir α de sorte que $\Delta < 0$: le champ se redresse par le théorème 1.

5.3. Etude du cas $a = 0$ et $b' \neq 0$

On étudie

$$\begin{cases} \dot{x} = bxy + cy^2 + dx + ey + f \\ \dot{y} = b'xy + c'y^2 + d'x + e'y + f'. \end{cases}$$

On pose $X = b'x - by$ et on obtient :

$$\begin{cases} \dot{X} = Cy^2 + DX + Ey + F \\ \dot{y} = Xy + C'y^2 + D'X + E'y + F'. \end{cases}$$

Si $C = 0$, par échange de X et y, on se ramène au cas $b' = c' = 0$, étudié en 5.5.

Si $C \neq 0$ et $D \neq 0$, on tire X de l'équation $\dot{X} = 0$ et on reporte dans $\dot{y} = 0$. On obtient une équation du troisième degré qui a une racine : il y a un point fixe.

Si $C \neq 0$ et $D = 0$, on s'est ramené à :

$$\begin{cases} \dot{x} = cy^2 + ey + f \\ \dot{y} = xy + c'y^2 + d'x + e'y + f'. \end{cases} \tag{$E0$}$$

Si $\dot{x} = 0$ n'a pas de racines, le champ se redresse par le théorème 2.

Si $\dot{x} = 0$ a une racine différente de $-d'$, il y a un point fixe.

Sinon le champ est :

$$\begin{cases} \dot{x} = c(y + d')^2 \\ \dot{y} = x(y + d') + c'(y + d')^2 + (e' - 2c'd')(y + d') + f' + c'd'^2 - e'd'. \end{cases}$$

On pose $X = x + e' - 2c'd'$, $Y = y + d'$ et on obtient un système du type :

$$\begin{cases} \dot{x} = cy^2 \\ \dot{y} = xy + c'y^2 + f''. \end{cases} \tag{$CP1$}$$

Il est clair que $f'' \neq 0$ car sinon il y a un point fixe. Si f'' n'est pas du signe de c on pose $Y = -y$ de sorte que cela le devienne.

5.4. Etude du cas $a = b' = 0$, $c' \neq 0$

On étudie

$$\begin{cases} \dot{x} = bxy + cy^2 + dx + ey + f \\ \dot{y} = c'y^2 + d'x + e'y + f'. \end{cases}$$

Si $b = 0$, on pose $X = y$ et $Y = c'x - cy$ et on se ramène au cas 5.5 : $b' = c' = 0$.

Si $bd' \neq 0$, on tire x de $\dot{y} = 0$, on reporte dans $\dot{x} = 0$ et on obtient une équation du troisième degré, donc un point fixe.

Sinon on a $d' = 0$, $b \neq 0$ et $c' \neq 0$.

Si $\dot{y} = 0$ n'a pas de racines, le champ se redresse (théorème 2, où l'on inverse les rôles de x et y).

Si $\dot{y} = 0$ a une racine différente de $-d/b$, il y a un point fixe.

Sinon il y a une racine double et par translation on se ramène à :

$$\begin{cases} \dot{x} = bxy + cy^2 + Ey + F \\ \dot{y} = c'y^2 \end{cases}$$

et ce champ se redresse par des droites $y = k$ avec $k \neq 0$, la preuve étant la même que celle du théorème 2.

5.5. Etude du cas $b' = c' = 0$

On considère

$$\begin{cases} \dot{x} = ax^2 + bxy + cy^2 + dx + ey + f \\ \dot{y} = d'x + e'y + f'. \end{cases}$$

Si $d' \neq 0$, on pose $X = d'x + e'y + f'$ et on est ramené à $\dot{y} = X$.

Si $d' = e' = f' = 0$, on a $\dot{y} = 0$.

Si $d' = 0$, $e' \neq 0$, on se ramène à $\dot{Y} = yY$ par $Y = e'y + f'$ et $T = e't$.

Si $d' = e' = 0$, $f' \neq 0$, on se ramène à $\dot{y} = 1$ par $T = f't$.

On va traiter ces quatre cas dans les sous-sections suivantes.

5.5.1. Premier cas

Etudions

$$\begin{cases} \dot{x} = ax^2 + bxy + cy^2 + dx + ey + f \\ \dot{y} = x. \end{cases}$$

Si $a = b = c = 0$, le champ est affine. Comme il n'y a pas de points fixes on a $e = 0$ et $f \neq 0$. En posant $X = x - dy$ on obtient $\dot{X} = f$ et le champ se redresse par le théorème 2.

Si $b = c = 0$, $a \neq 0$, on a $e = 0$, $f \neq 0$ (pas de point fixe). Dans ce cas, si $ax^2 + dx + f$ n'a pas de racines, le champ se redresse (théorème 2); sinon, on a $\dot{x} = a(x - A)(x - B)$, avec $AB \neq 0$. Si $AB > 0$, le champ se redresse, car il suffit de poser $X = x + a(A + B)y$, pour obtenir $\dot{X} = a(x^2 + AB)$ et conclure par le théorème 2. Sinon on pose $X = ax$ et $Y = ay$ et on se ramène à

$$\begin{cases} \dot{x} = (x - a)(x - b) \\ \dot{y} = x \\ ab < 0. \end{cases} \qquad (CP2a)$$

Si $c = 0$, $b \neq 0$, on a $e = 0$ et $f \neq 0$ (pas de point fixe) et on se ramène, par $X = \lambda x$, $Y = \mu(y + \frac{d}{b})$ $T = \nu t$, avec $\lambda^3 = bf^{-2}$, $\mu = \lambda^2 f$ et $\nu = \lambda f$, à :

$$\begin{cases} \dot{x} = ax^2 + xy + 1 \\ \dot{y} = x. \end{cases} \qquad (CP2b)$$

Sinon on a $c \neq 0$. Si $e^2 - 4cf \geq 0$ il y a des points fixes. Supposons $e^2 - 4cf < 0$. Si $b^2 - 4ac < 0$, le champ se redresse par le théorème 1. Si $b^2 - 4ac = 0$, le champ se redresse par le théorème 2 en posant $X = cx + (be/2 - cd)y$, ce qui donne

$$\begin{cases} \dot{X} = (\frac{bX}{2} + Ay + \frac{e}{2})^2 - \frac{e^2 - 4cf}{4} \\ \dot{Y} = \frac{X}{c} - (\frac{be}{2c} - d)Y. \end{cases}$$

Si $b^2 - 4ac > 0$, on pose $\alpha = \dfrac{be - 2cd}{b^2 - 4ac}$, $\beta = \dfrac{bd - 2ae}{b^2 - 4ac}$, $\gamma = \dfrac{(b^2 - 4ac)(4cf - e^2) + (eb - 2cd)^2}{4c(b^2 - 4ac)}$. On peut choisir $\lambda^4 = c\gamma^{-3}$, $\mu = \lambda^2\gamma$, $\nu = \mu/\lambda$ et poser $X = \lambda(x + \alpha)$, $Y = \mu(y + \beta)$, $T = \nu t$, ce qui donne

$$\begin{cases} \dot{X} = (AX + Y)(BX + Y) + 1 \\ \dot{Y} = X + C. \end{cases} \qquad (CP2c)$$

5.5.2. Deuxième cas

Etudions

$$\begin{cases} \dot{x} = ax^2 + bxy + cy^2 + dx + ey + f \\ \dot{y} = 0. \end{cases}$$

Si $\dot{x} = 0$ a une racine, il y a un point fixe. Sinon le champ se redresse par le théorème 2.

5.5.3. Troisième cas

Etudions

$$\begin{cases} \dot{x} = ax^2 + bxy + cy^2 + dx + ey + f \\ \dot{y} = y. \end{cases}$$

Notons que si $b^2 - 4ac < 0$ le champ se redresse (théorème 1). On suppose dans la suite que $b^2 - 4ac \geq 0$. On suppose également que $f \neq 0$, car sinon il y a un point fixe.

** Si $a = b = 0$ il faut $d = 0$, sinon il y a un point fixe. On est donc ramené à

$$\begin{cases} \dot{x} = cy^2 + ey + f \\ \dot{y} = y. \end{cases}$$

Si $\dot{x} = 0$ n'a pas de racines, le champ se redresse (théorème 2). Si $c = 0$ ou $cf > 0$, en posant $X = x - ey$, on a $\dot{X} = cy^2 + f$ et le champ se redresse par le théorème 2. Dans le cas $cf < 0$ on pose $X = \lambda y$, $Y = \mu(x - ey)$ avec $\mu = -1/f$, $\lambda = \sqrt{-c/f}$ et on est amené à

$$\begin{cases} \dot{X} = X \\ \dot{Y} = (X-1)(X+1). \end{cases} \qquad (CP3)$$

** Si $a = 0$, $b \neq 0$, il faut aussi $d = 0$. Dans ce cas on pose

$$X = by \quad Y = \frac{1}{f}[x + \frac{c}{b}y + \frac{be+c}{b^2}],$$

et on obtient :

$$\begin{cases} \dot{X} = X \\ \dot{Y} = XY + 1. \end{cases} \qquad (CP4)$$

** Sinon on a $a \neq 0$. Dans ce cas on pose $X = y$ et $Y = x + \alpha y$, où α vérifie $a\alpha^2 - b\alpha + c = 0$ (par hypothèse $b^2 - 4ac \geq 0$ et $a \neq 0$), et on obtient :

$$\begin{cases} \dot{X} = X \\ \dot{Y} = B'XY + aY^2 + D'X + E'Y + F'. \end{cases}$$

Dans ce système, si $B' \neq 0$, on peut se ramener à $D' = 0$, $B' = a = 1$, par $X = B'x$, $Y = a(y + \frac{D'}{B'})$ d'où :

$$\begin{cases} \dot{x} = x \\ \dot{y} = xy + y^2 + Ay + B. \end{cases} \qquad (CP5)$$

Si$B' = 0$ on pose $Y = y - D'x + \frac{E'}{2a}$. Alors $\dot{Y} = a[(Y + D'x)^2 + A]$ et le champ se redresse si $A > 0$ (théorème 2), sinon il a un point fixe.

5.5.4. Dernier cas

Etudions

$$\begin{cases} \dot{x} = ax^2 + bxy + cy^2 + dx + ey + f \\ \dot{y} = 1. \end{cases}$$

Si $a = 0$, le champ se redresse par le théorème 2 en échangeant x et y.

Si $a \neq 0$, on pose $X = ax + by/2 + d/2$. Alors $B = D = 0$, $A = 1$ donc on se ramène à

$$\begin{cases} \dot{x} = x^2 + Cy^2 + Ey + F \\ \dot{y} = 1. \end{cases}$$

Si $C > 0$, le champ se redresse par le théorème 1.

Si $C = 0$, $E = 0$, le champ se redresse car, en posant $Y = x + (1 + |F| - F)y$, on a $\dot{Y} = x^2 + 1 + |F|$, et donc le théorème 2 s'applique (en échangeant les rôles de x et y).

Si $C = 0$, $E \neq 0$, en posant $X = \lambda x$, $Y = \mu(Ey + F)$ et $T = \nu t$ avec $\mu^3 = -E^{-2}$, $\lambda = (\mu E)^{-1}$ et $\nu = \mu E$ on obtient :

$$\begin{cases} \dot{x} = x^2 - y \\ \dot{y} = 1. \end{cases} \qquad (CP6)$$

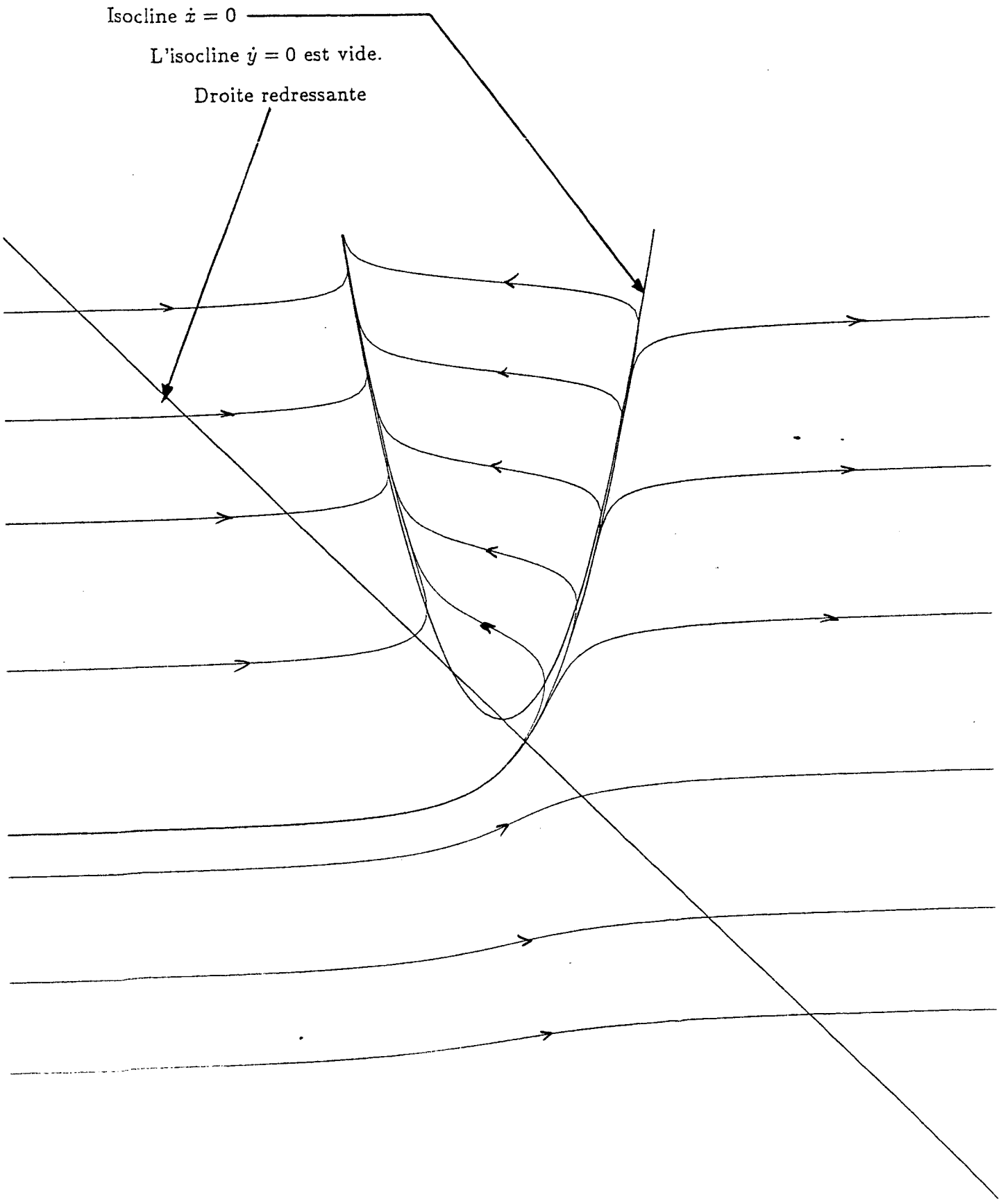
Si $C < 0$, on pose $X = \lambda x$, $Y = \mu(y + \frac{E}{2C})$, et $T = \nu t$ avec $\mu^4 = -C$, $\lambda = 1/\mu$, $\nu = \mu$, d'où :

$$\begin{cases} \dot{x} = x^2 - y^2 + a \\ \dot{y} = 1. \end{cases} \qquad (CP7)$$

On va étudier dans la suite les cas (CPi) pour i=0,1,...,7.

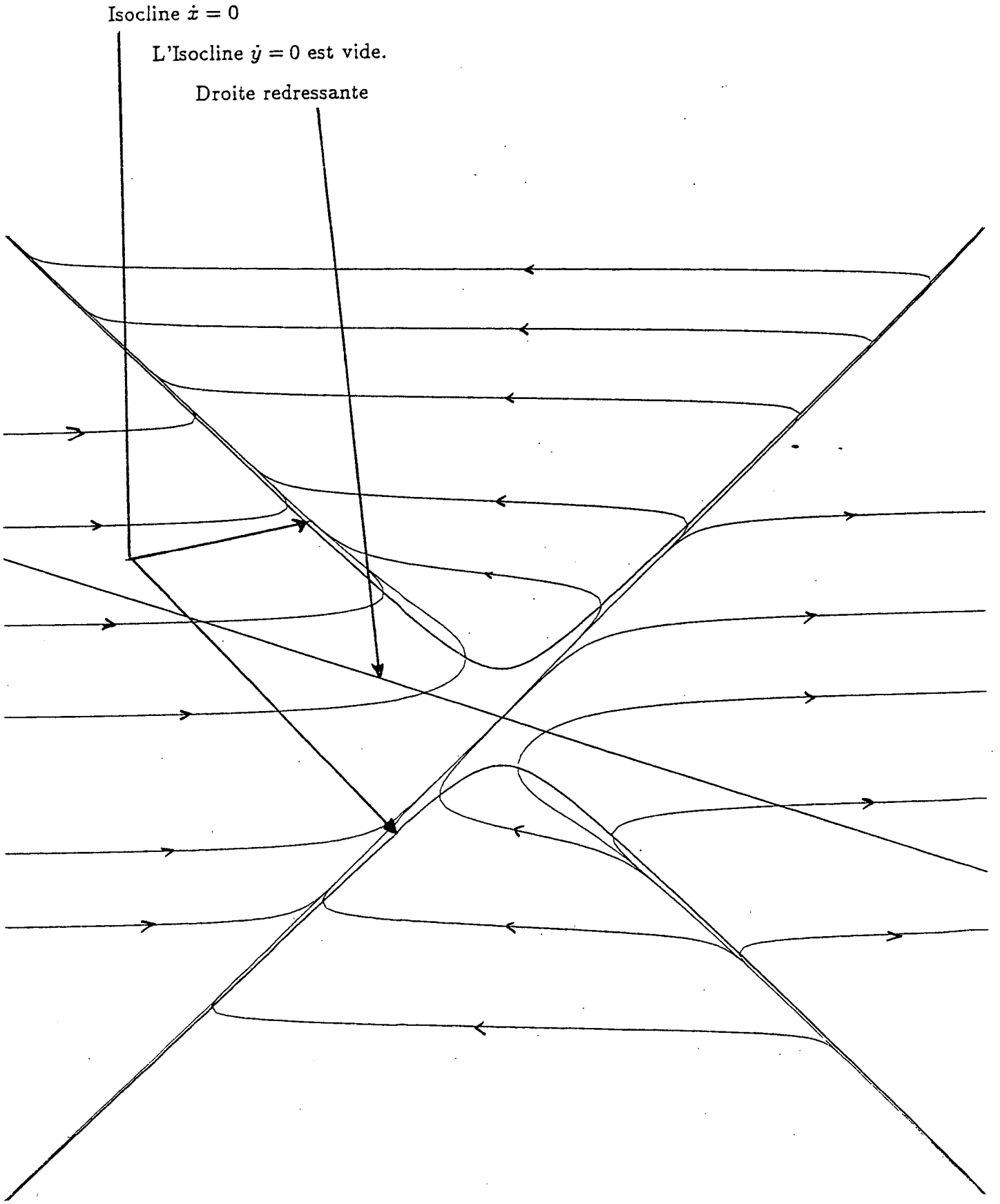
Cas CP6

$$\begin{cases} \dot{x} = x^2 - y \\ \dot{y} = 1 \end{cases}$$



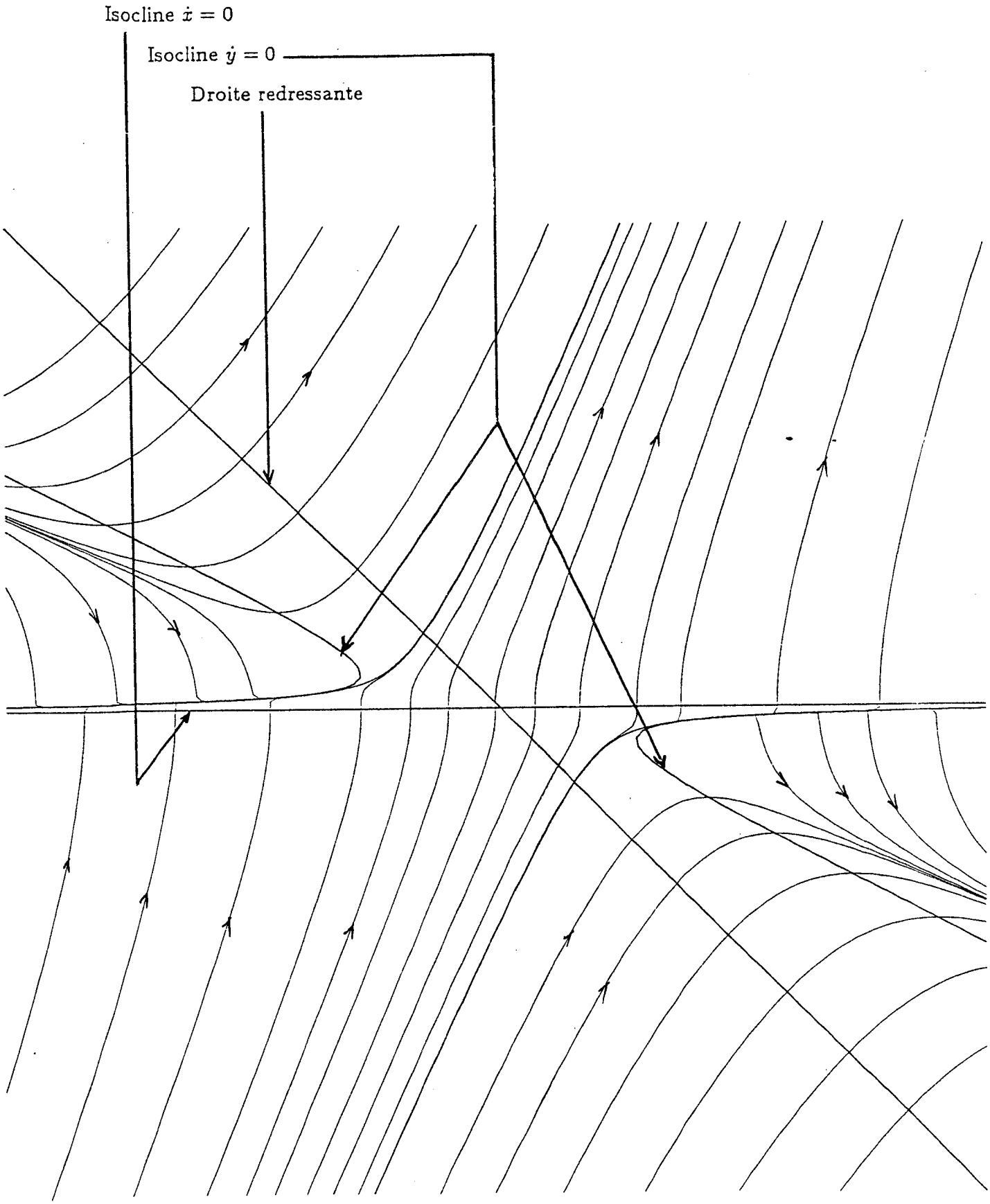
Cas CP7

$$\begin{cases} \dot{x} = x^2 - y^2 + 1 \\ \dot{y} = 1 \end{cases}$$



Cas CP1

$$\begin{cases} \dot{x} = y^2 \\ \dot{y} = xy + 2y^2 + 1 \end{cases}$$



6. Etude de CP6

$$\begin{cases} \dot{x} = x^2 - y \\ \dot{y} = 1 \end{cases} \qquad (CP6)$$

On pose $u = x + y$, et le champ se redresse par $u = 0$.

En effet $\dot{u} = x^2 + x + 1 - u$, donc si $u = 0$, $\dot{u} > 0$, et la droite est transverse au champ.

Si $u \leq 0$, alors $\dot{u} \geq \frac{3}{4}$. Si u reste constamment ≤ 0 pour $t > 0$, on en déduit que T_+ est fini, ce qui est absurde, car y et u resteraient bornés, d'où si $t \to T_+$, u devient positif.

Si $u_0 > 0$, on regarde ce qui se passe lorsque $t \to T_-$. Si $T_- = -\infty$ alors $y \to -\infty$, $x \to -\infty$, donc $u = x + y$ devient négatif. Sinon, comme y est borné, c'est que x diverge, donc $x \to -\infty$, et u aussi.

7. Etude de CP7

$$\begin{cases} \dot{x} = x^2 - y^2 + a \\ \dot{y} = 1 \end{cases} \qquad (CP7)$$

Soit $b = 2 + |a|$ et $u = x + by$. Le champ se redresse par $u = 0$.

En effet $\dot{u} = x^2 - y^2 + a + b$. Si $u = 0$, $\dot{u} = (b^2 - 1)y^2 + a + b > 0$: le champ est transverse à $u = 0$.

Si $u_0 < 0$ on regarde ce qui se passe si $t \to T_+$.

Si T_+ est fini alors y est borné, $x \to +\infty$ et $u \to +\infty$.

Si $T_+ = +\infty$ et qu'on suppose u constamment < 0, alors $y \to +\infty$ et $x < -2y$, d'où $x^2 - y^2 \geq 3y^2$ et $\dot{u} \to +\infty$, d'où par le lemme 3, $u \to +\infty$, ce qui est absurde. On fait de même pour $u_0 > 0$, en remarquant que le champ est pair et la droite symétrique par rapport à l'origine.

8. Etude de CP1

$$\begin{cases} \dot{x} = ay^2 \\ \dot{y} = xy + by^2 + c \end{cases} \qquad (CP1)$$

On sait que $ac > 0$.

On choisit $\alpha > 0$ petit, de sorte que $a + \alpha b$ et $a + \alpha b - \alpha^2$ soient du signe de a.

On pose $u = x + \alpha y$ et on va montrer par $u = 0$ redresse le champ.

On a :

$$\dot{u} = (a + \alpha b - \alpha^2)y^2 + \alpha c + \alpha u y = (a + \alpha b)y^2 + \alpha xy + \alpha c$$

La première égalité montre que la droite est transverse au champ.

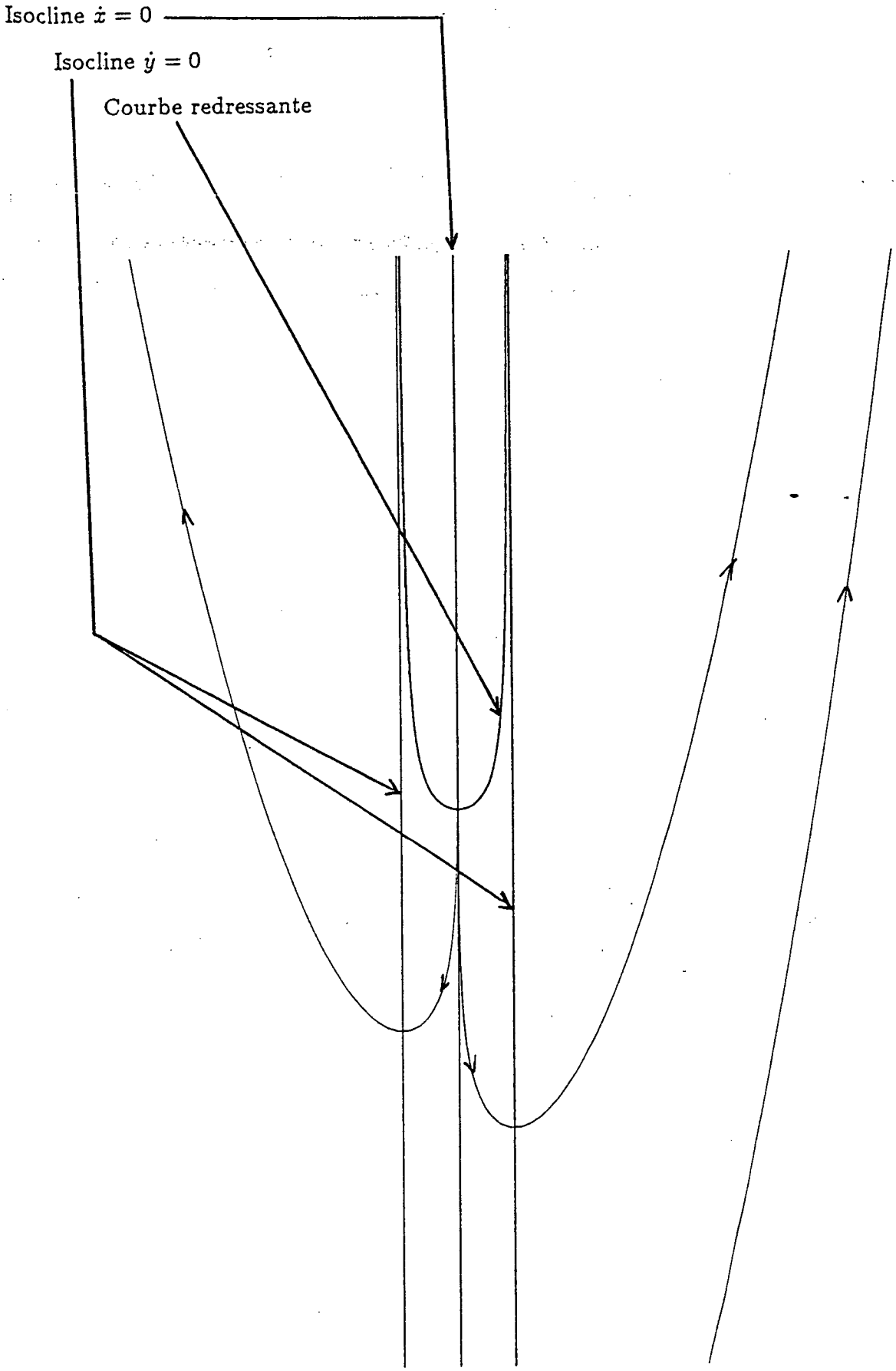
Notons que $a \neq 0$ entraîne, par le lemme 3, que x ne peut être borné si $t \to T_\pm$, et l'hypothèse sur α entraîne la même conclusion pour u.

Supposons $a > 0$. Dans ce cas, si pour tout $t > 0$ on a $u < 0$, si $t \to T_+$, $x \to +\infty$, y devient négatif, la première formule pour $\dot{u}$ montre que $\dot{u} > 0$, et donc u est borné, absurde. Par conséquent, si on part de $u_0 < 0$, on coupe la droite $u = 0$ pour un certain instant $t > 0$. De même, si on part de $u_0 > 0$, on coupe la droite pour un certain $t < 0$, car le champ et la droite sont invariants par la transformation $x = -x$, $y = -y$, $t = -t$.

Supposons $a < 0$. La preuve se fait de la même manière, sachant que si pour tout $t > 0$ on avait $u > 0$, on aurait $x \to -\infty$ d'où $y \to +\infty$, et la deuxième formule pour $\dot{u}$ montre maintenant que $\dot{u} < 0$, donc u borné, absurde.

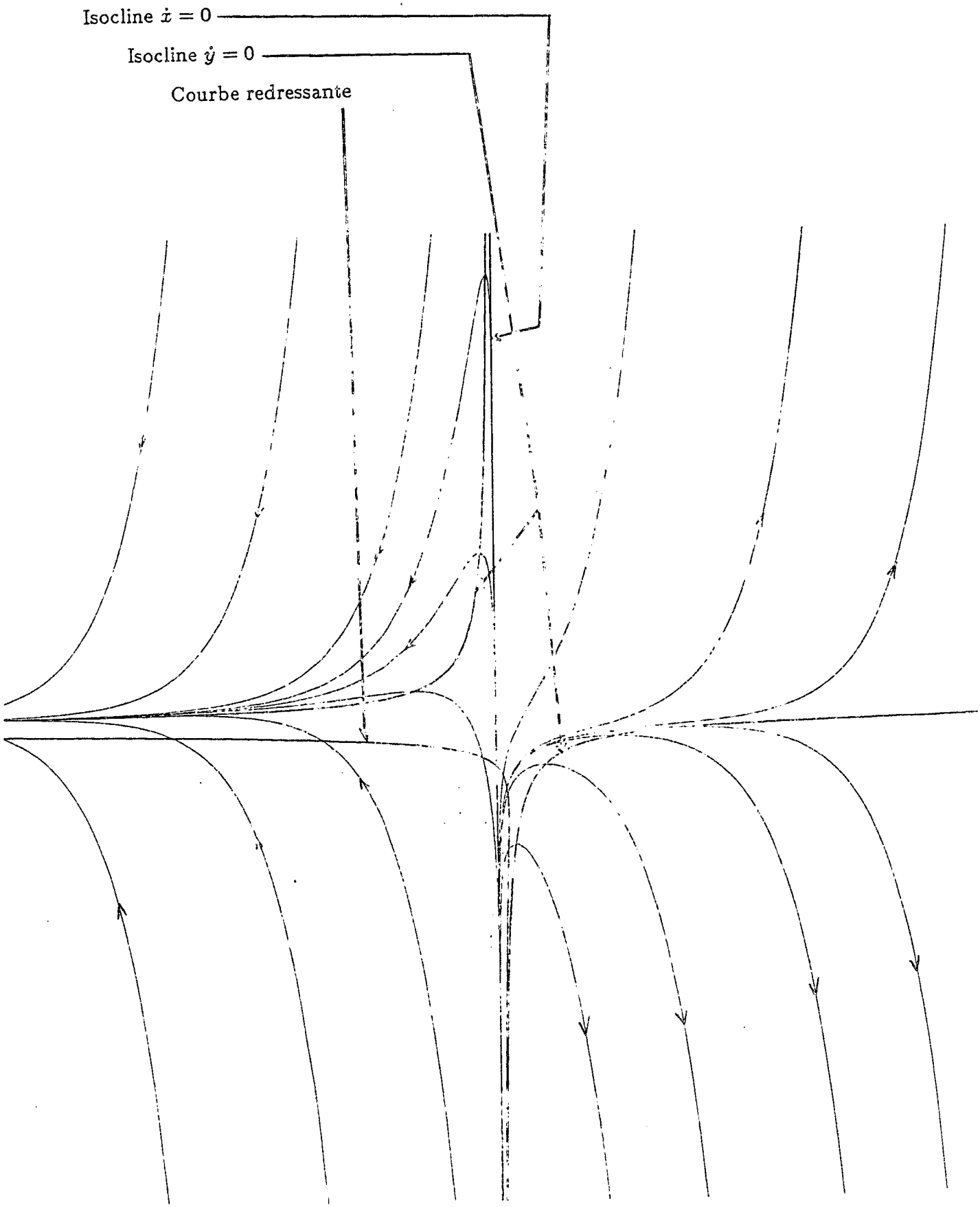
Cas CP3

$$\begin{cases} \dot{x} = x \\ \dot{y} = x^2 - 1 \end{cases}$$



Cas CP4

$$\begin{cases} \dot{x} = x \\ \dot{y} = xy + 1 \end{cases}$$



9. Etude de CP3

$$\begin{cases} \dot{x} = x \\ \dot{y} = (x-1)(x+1) \end{cases} \qquad (CP3)$$

Ce champ se redresse mais pas par des droites.

Notons que ce champ s'intègre explicitement et on a $x = x_0 e^t$ et $y = \dfrac{x_0^2(e^{2t}-1)}{2} - t + y_0$, de sorte que $T_- = -\infty$ et $T_+ = +\infty$.

Montrons d'abord qu'une droite ne peut pas redresser le champ.

Les droites $x = \gamma$ ne redressent pas le champ car $x = 0$ est une trajectoire.

Sur la droite $u = y - \alpha x - \beta = 0$, on a $\dot{u} = x^2 - \alpha x - 1$, qui n'est pas de signe constant, et la droite ne peut donc redresser le champ.

On prend la courbe

$$\Gamma : y = f(x) = \frac{x^2}{1-x^2} \qquad -1 < x < 1$$

et le domaine D dont le bord est Γ

$$D = \{(x,y) | \quad -1 < x < 1,\ y \geq f(x)\}.$$

La courbe Γ redresse le champ. En effet, toute trajectoire de D quitte D si $t \to T_+$, car ceci est clair pour la trajectoire $x = 0$, tandis que $|x| \to +\infty$ le long des autres.

Toute trajectoire en dehors de D y rentre si $t \to T_-$, car dans ce cas $x \to 0$ et $y \to +\infty$.

Le champ est transverse sur Γ. En effet la tangente à Γ a pour coordonnées $(1, f'(x))$. Le déterminant de cette tangente et du champ est $\dot{y} - \dot{x}f'(x)$ mais sur Γ, $\dot{y} < 0$ et $\dot{x}f'(x) = xf'(x) \geq 0$, d'où $\dot{y} - \dot{x}f'(x) < 0$.

10. Etude de CP4

$$\begin{cases} \dot{x} = x \\ \dot{y} = xy + 1 \end{cases} \qquad (CP4)$$

Ce champ se redresse mais pas par des droites.

Si $\dot{y} = 0$, alors $\ddot{y} = \dot{x}y = xy = -1$. Ceci montre que $\dot{y}$ s'annule une fois au plus le long de la trajectoire, et ce point correspondant est un maximum absolu.

Les droites $x = \gamma$ ne redressent pas le champ car $x = 0$ est une trajectoire.

Soit $y = \alpha x + \beta$ une droite.

** Si $\alpha \geq 0$, on choisit y_0 avec $y_0 < 0$ et $y_0 < \beta$. On pose $x_0 = -1/y_0$, de telle sorte que $\dot{y}$ s'annule en ce point, et donc $y \leq y_0$ et x reste positif.

Donc $y - \alpha x - \beta$ est toujours négatif et la trajectoire issue de (x_0, y_0) ne peut couper la droite.

** Si $\alpha < 0$, on pose $u = y - \alpha x - \beta$. Alors $\dot{u} = xy + 1 - \alpha x$ et si $u = 0$, $\dot{u} = \alpha x^2 + (\beta - \alpha)x + 1$: le champ n'est pas partout transverse à la droite, car cette expression n'est pas de signe constant.

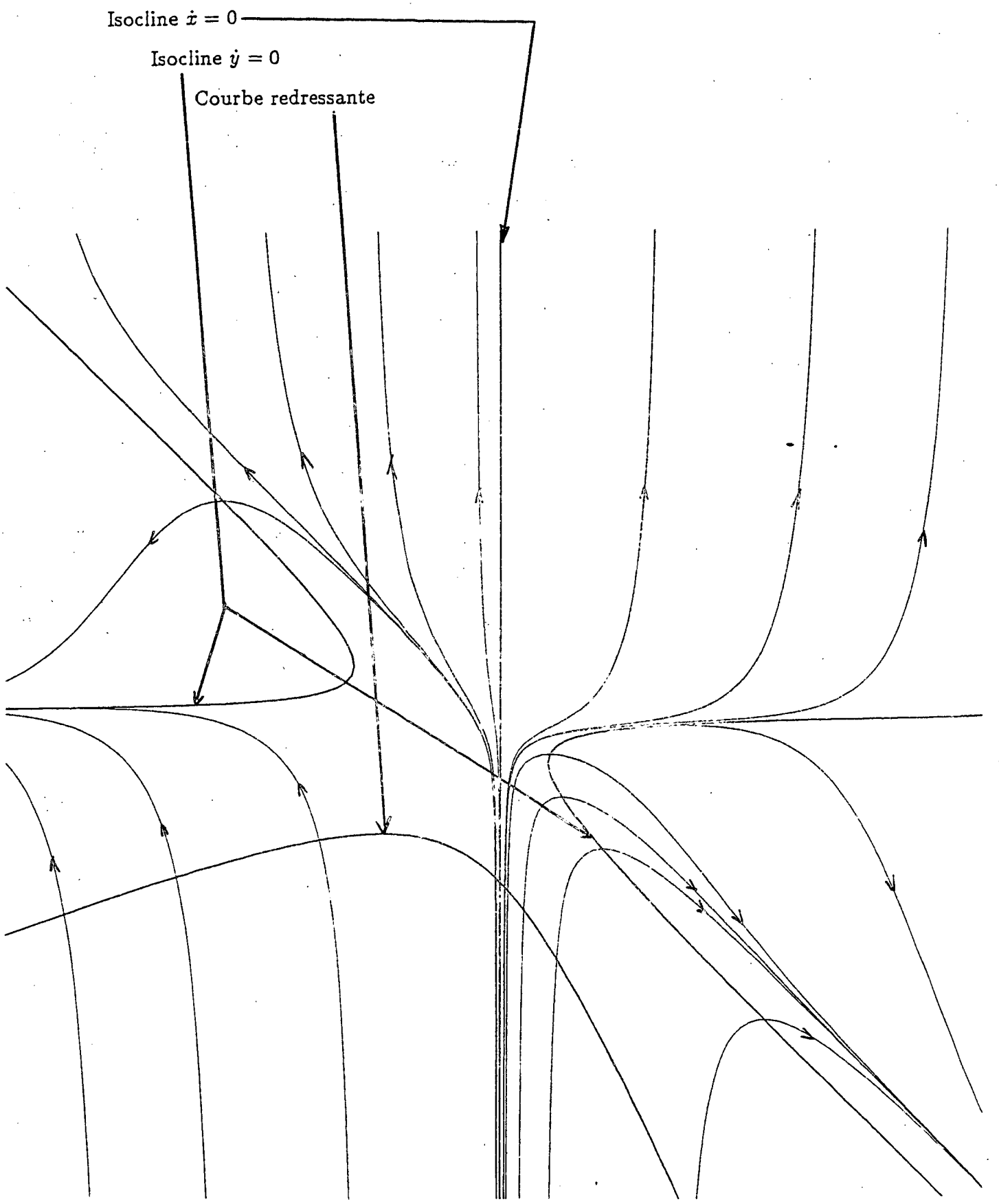
Donc le champ ne se redresse pas par des droites.

On considère la courbe Γ d'équations

$$\Gamma : x = -\frac{1+y}{y^2} \qquad y < 0$$

Cas CP5

$$\begin{cases} \dot{x} = x \\ \dot{y} = xy + y^2 + y + 1 \end{cases}$$



et le domaine D dont le bord est Γ

$$D = \{(x,y)| \quad x + \frac{1+y}{y^2} < 0, \quad y < 0\}.$$

Cette courbe redresse le champ.

En effet, posant $w = x + \frac{1+y}{y^2}$, on a pour $w = 0$, $\dot{w} = \frac{1+(1-y)(1+y)^2}{y^4}$ qui est > 0, car $y < 0$. Le champ est donc transverse à la courbe.

Notons que pour toute trajectoire du champ on a $T_- = -\infty$ et $T_+ = +\infty$, et que les trajectoires du champ sont de la forme

$$x = x_0 e^t \quad y = y_0 e^{x_0(e^t-1)} + \int_0^t e^{x_0(e^t-e^s)} ds.$$

La trajectoire $x = 0$ coupe Γ. Si $x_0 \neq 0$ on a $y = y_0 e^{x-x_0} + e^x \int_{Log|x_0|}^{t+Log|x_0|} e^{-\epsilon e^v} dv$ où ϵ est le signe de x_0. Si $t \to -\infty$, e^x tend vers 1, si s tend vers $-\infty$ l'intégrande équivaut à 1, donc $y \sim t$, $w \sim x_0 e^t + 1/t$ donc devient < 0, et la trajectoire rentre dans D.

Faisons tendre t vers $+\infty$. Si $x_0 > 0$ alors $x \to +\infty$ et la trajectoire sort de D. Si par contre $x_0 < 0$, alors $x \to -\infty$, il suffit de prouver que y devienne > 0 pour montrer que la trajectoire sort de D. Si cela n'était pas le cas, comme $\dot{y} > 1$, y serait borné, ce qui est absurde, car $T_+ = +\infty$.

11. Etude de CP5

$$\begin{cases} \dot{x} = x \\ \dot{y} = xy + y^2 + ay + b \end{cases} \qquad (CP5)$$

Ce champ se redresse mais pas par des droites.

En effet

** Les droites $x = \gamma$ ne redressent pas le champ car $x = 0$ est une trajectoire.

** Comme précédemment, les extrema de y sont des maxima absolus car si $\dot{y} = 0$, $\ddot{y} = xy = -(y^2 + ay + b) < 0$ car il n'y a pas de points fixes.

**si $u = y - \alpha x - \beta$ et $\alpha \geq 0$, on choisit $y_0 < 0$, $y_0 < \beta$ et $x_0 = -\frac{y_0^2 + ay_0 + b}{y_0} > 0$; sur la trajectoire on a alors $\forall t$ $y \leq y_0$, $x > 0$ et $u < 0$.

** si $\alpha < 0$, on prend $y_0 > 0$ et $x_0 = -\frac{y_0^2 + ay_0 + b}{y_0} < 0$, y_0 petit de sorte que $x_0 < 0$ et $u_0 < 0$ (ce qui est possible car $b > 0$ donc si $y_0 \to 0$, $x_0 \to -\infty$). Si $t > 0$, y et x décroissent, donc u aussi et u reste < 0. Si $t \to T_-$, $y \to -\infty$ et x est borné donc $u \to -\infty$.

Donc aucune droite ne redresse le champ.

Soit $B = 4(|a|+1)^2$, Γ la courbe

$$\Gamma : y = -\sqrt{2x^2 + B}$$

et D le domaine dont le bord est Γ

$$D = \{(x,y)| \quad y \leq -\sqrt{2x^2 + B}\}.$$

Dans D on a $\dot{y} > 0$ car

$$\dot{y} = xy + \frac{y^2}{2} + \frac{y^2}{4} + ay + b + \frac{y^2}{4} \geq xy + \frac{y^2}{2} + \frac{y^2}{4} + ay + b + \frac{x^2}{2} + \frac{B}{4} = \frac{(x+y)^2}{2} + \frac{(y+2a)^2}{4} + \frac{B - 4a^2 + 4b}{4}$$

et $b > 0$, sinon il y a des points fixes. Comme $D \cap \{(x,y)| \quad y \geq k\}$ est compact, toute trajectoire de D sort de D si $t \to T_+$.

De plus si $t \to T_-$, x est borné, donc $|y| \to \infty$, en fait $y \to -\infty$, et donc toute trajectoire rentre dans D.

On vérifie enfin que Γ est bien transverse au champ, puisque $\dot{y} + \frac{2x^2}{\sqrt{2x^2+B}} > 0$ sur Γ.

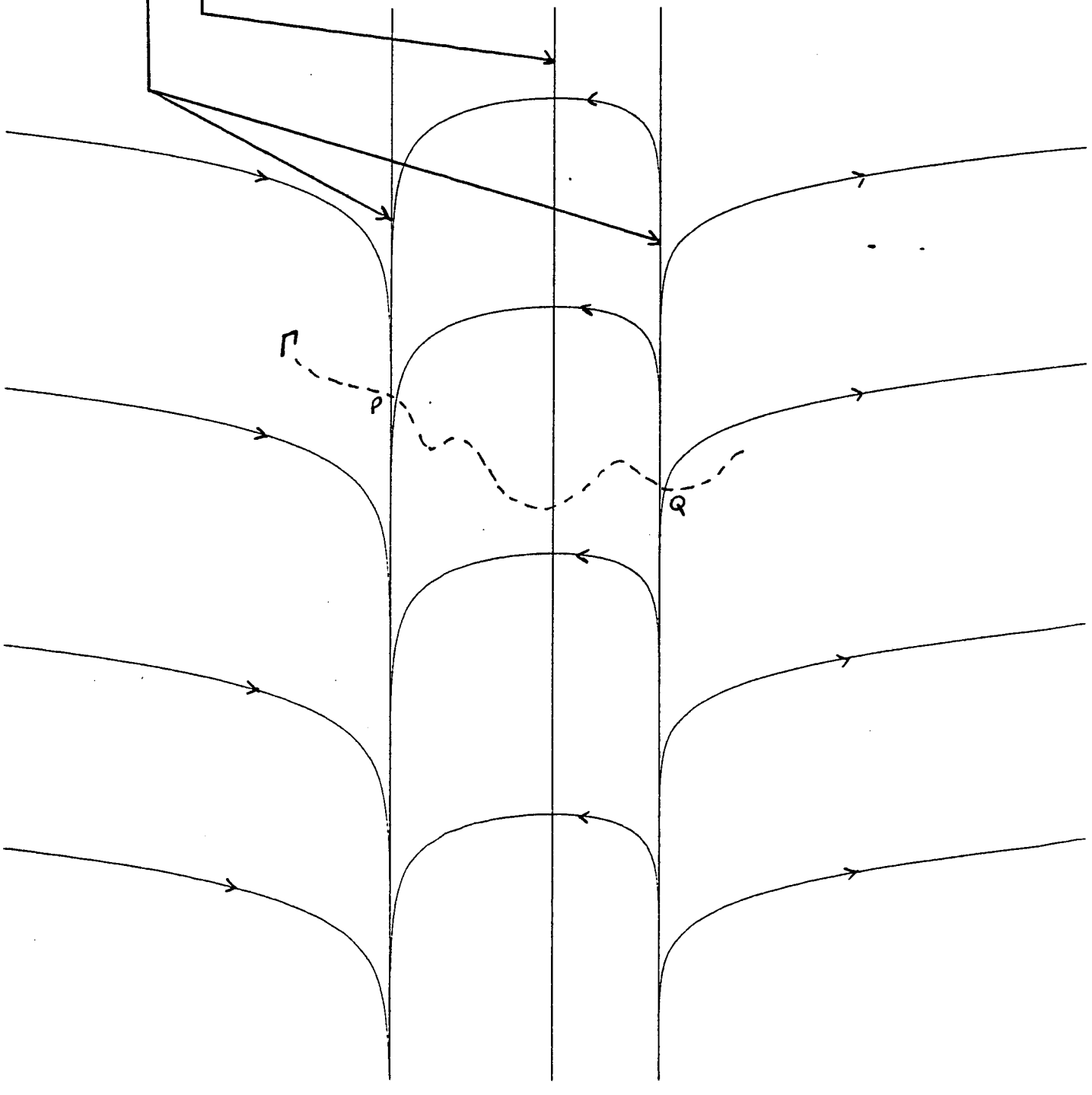
Cas CP2a

$$\begin{cases} \dot{x} = (x-2)(x-3) \\ \dot{y} = x \end{cases}$$

Isocline $\dot{x} = 0$

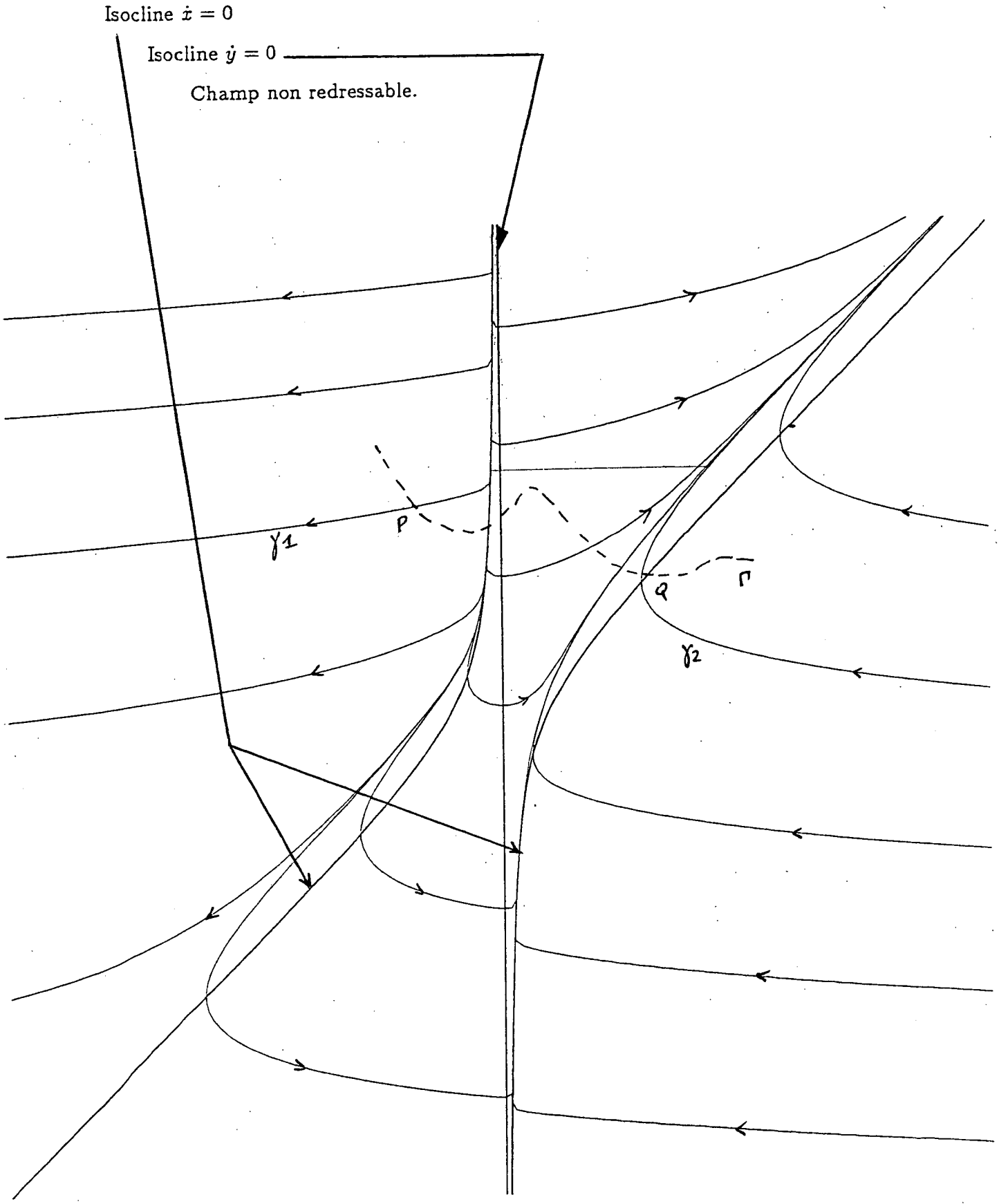
Isocline $\dot{y} = 0$

Champ non redressable.



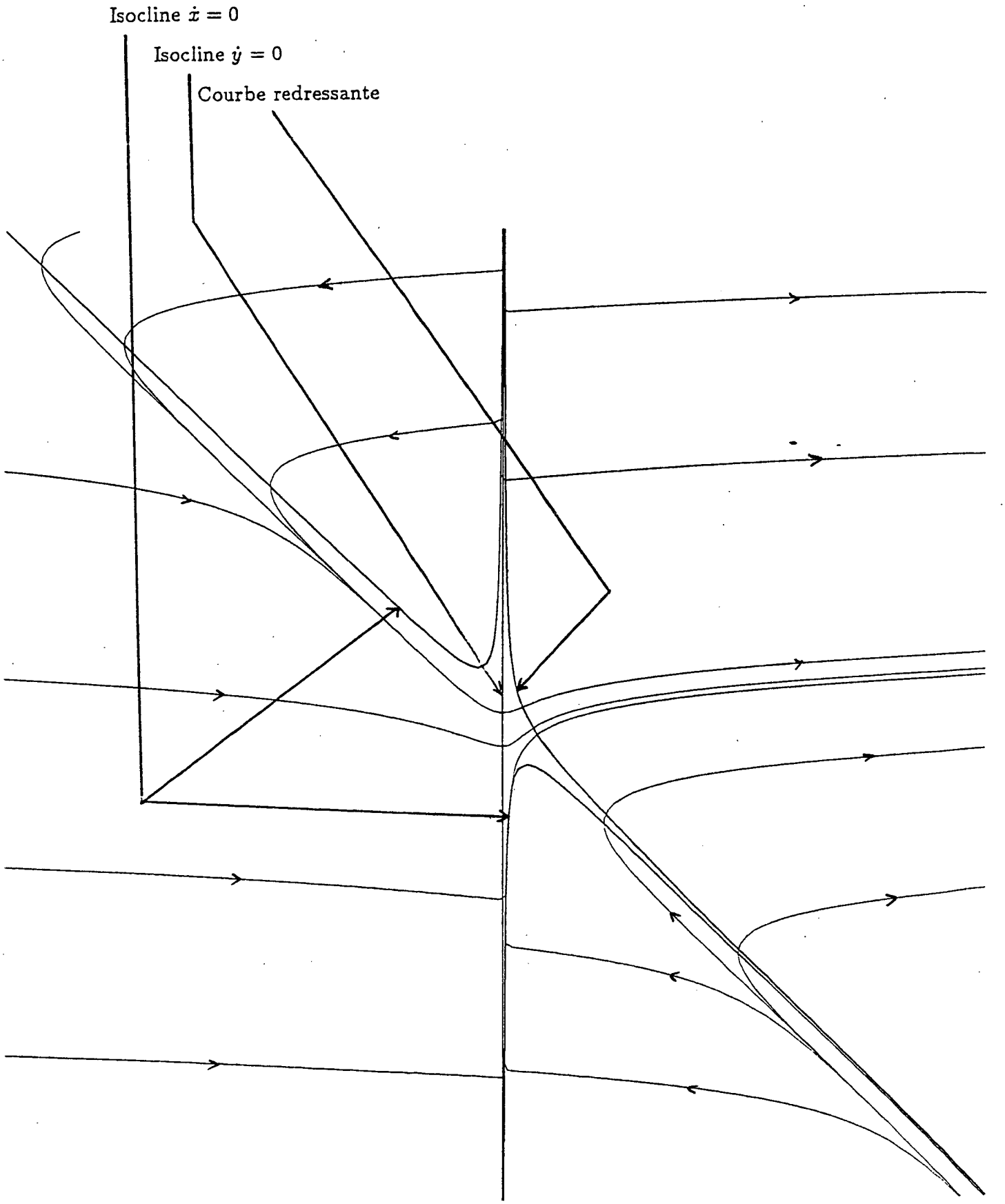
Cas CP2b

$$\begin{cases} \dot{x} = -x^2 + xy + 1 \\ \dot{y} = x \end{cases}$$



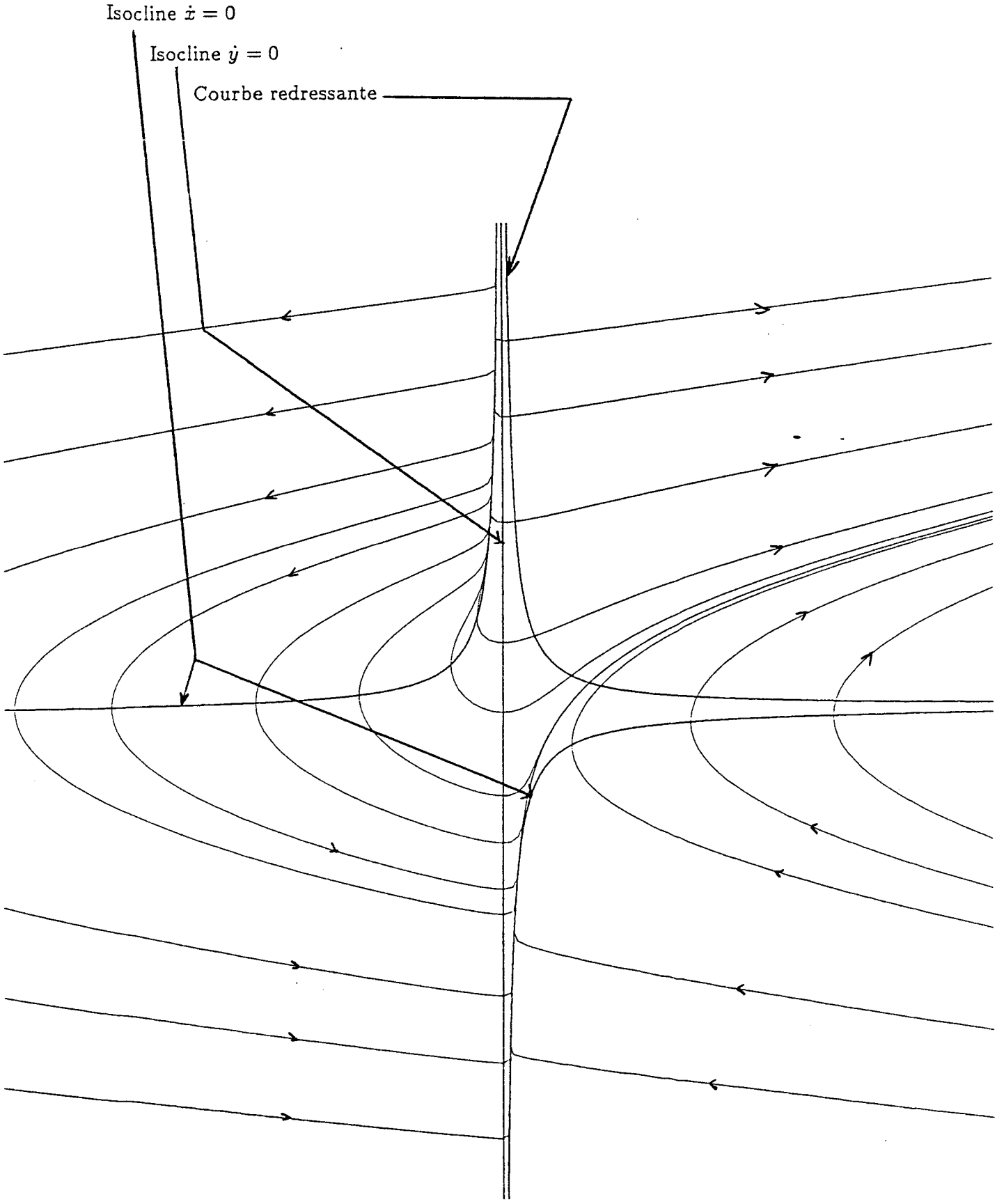
Cas CP2b

$$\begin{cases} \dot{x} = x^2 + xy + 1 \\ \dot{y} = x \end{cases}$$



Cas CP2b

$$\begin{cases} \dot{x} = xy + 1 \\ \dot{y} = x \end{cases}$$



12. Etude de CP2

12.1. Etude de CP2a

$$\begin{cases} \dot{x} = (x-a)(x-b) \\ \dot{y} = x \\ ab < 0 \end{cases} \qquad (CP2a)$$

Le champ ne se redresse pas.

En effet soit Γ une courbe redressante. Γ coupe les trajectoires $x = a$ et $x = b$ en deux points P et Q. L'arc de Γ reliant P à Q a un minimum en y disons y_0.

Soit D le domaine $a \leq x \leq b$, $y \leq y_0 - 1$. Alors Γ n'a pas de point d'intersection avec D.

Comme la trajectoire qui passe par $(0, y_0 - 2)$ est dans D, on a trouvé une trajectoire qui ne coupe pas la courbe redressante. Absurde.

12.2. Etude de CP2b

$$\begin{cases} \dot{x} = ax^2 + xy + 1 \\ \dot{y} = x \end{cases} \qquad (CP2b)$$

12.2.1. Cas $a < 0$

Le champ ne se redresse pas.

Posons $u = y + ax$. Alors $\dot{u} = x + aux + a$ et $\dot{x} = ux + 1$.

Il est facile de voir que les points critiques de x, y et u sont des minima absolus le long d'une trajectoire, car si $\dot{y} = 0$, $\ddot{y} = 1$, si $\dot{x} = 0$, $\ddot{x} = x^2 > 0$ et si $\dot{u} = 0$, $\ddot{u} = 1$.

Soit $x_0 = -1$ et $u_0 = -\dfrac{a-1}{a} > 0$. Alors u admet un minimum en 0 donc $\dot{u} > 0$ pour $t > 0$. Comme $\dot{u} = x + a\dot{x}$ on en déduit, pour $t > 0$, $\dot{x} < 0$ et $x < 0$.

Pour $t < 0$ on a $\dot{x} < 0$, sinon x aurait un maximum quelque part (initialement on a $\dot{x} < 0$). Donc $ux + 1 < 0$ et $x < 0$ car $u > 0$.

Soit γ_1 cette trajectoire sur laquelle $x < 0$. Soit γ_2 une trajectoire sur laquelle $x > 0$ (il suffit de prendre $x_0 = 1$ et $u_0 = -1$ comme condition initiale).

Soit Γ une coube redressante. Elle coupe γ_1 en P et γ_2 en Q. Γ a un maximum en y entre P et Q, disons y_0. Observons que y est monotone sur γ_1 et γ_2. Soit C_1 l'arc de γ_1 défini par $y \geq y_P$, C_2 l'arc de γ_2 défini par $y \geq y_Q$, C_3 l'arc de Γ compris entre P et Q, et D le domaine délimité par C_1, C_3, C_2 et contenant les points $(0, y)$ pour y assez grand. Alors Γ est extérieure à D, et la trajectoire qui passe par $(0, y_0 + 2)$ est intérieure à D donc ne coupe pas Γ.

12.2.2. Etude de $a \geq 0$

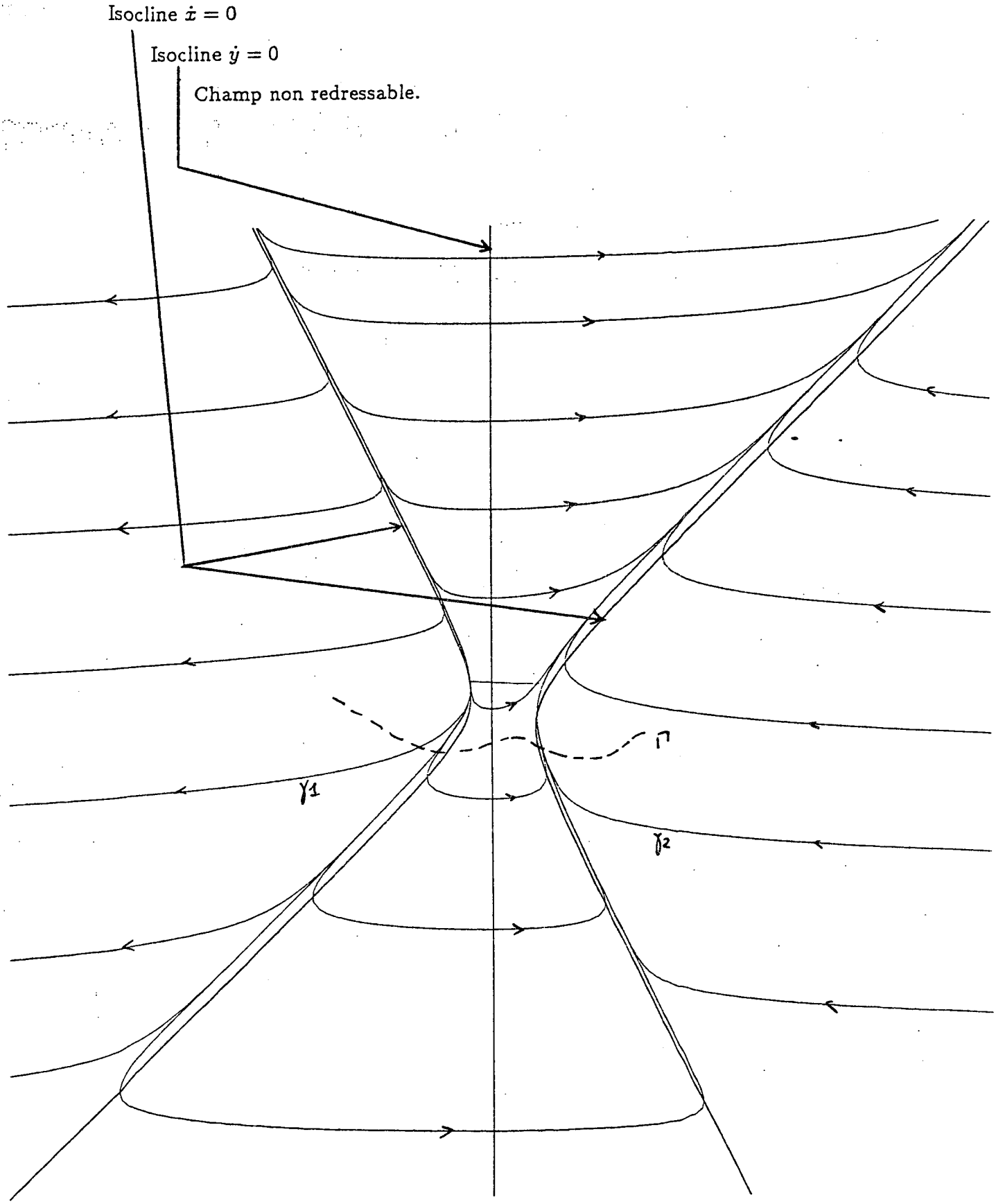
Ce champ se redresse mais pas par des droites.

Soit une droite $u = x - \alpha y - \beta$. Notons que les points critiques de x et y sont des minima absolus, car le calcul de la sous-section précédente ne faisait pas d'hypothèses sur le signe de a.

Premier cas $\alpha < 0$. On prend comme condition initiale $x_0 = -1$ et $y_0 = a + 1$. Alors $x \geq -1$ pour tout t. Si $t \to T_+$ et que x était toujours < 0, x serait borné, donc $y \to -\infty$ et $\dot{x} > 1$, et donc $T_+ < +\infty$, mais l'équation $\dot{y} = x$ entrainerait alors y borné. Par conséquent x devient positif, et le reste, parce que la droite $x = 0$ est transverse au champ. Donc x et y croissent, l'un au moins des deux tend vers $+\infty$, donc $u \to +\infty$. Regardons ce qui se passe si $t \to T_-$. On a toujours $x < 0$, car sinon on passerait par $x = 0$, et dans ce cas on aurait $\dot{x} = 1$, mais par hypothèse $\dot{x} < 0$ pour $t < 0$. Donc x est borné, par conséquent $y \to +\infty$, et $u \to +\infty$. La droite ne peut pas redresser le champ.

Cas CP2c

$$\begin{cases} \dot{x} = (-x+y)(2x+y)+1 \\ \dot{y} = x + 1/4 \end{cases}$$



On suppose maintenant $\alpha \geq 0$.

Prenons $x_0 > 0$ et $y_0 = -\dfrac{ax_0^2+1}{x_0}$. On prend (x_0, y_0) comme condition initiale. On choisit x_0 assez grand pour que u_0 soit > 0. Alors pour tout t, $x \geq x_0$. Donc $\dot{y} > 0$ et donc si $t < 0$, $y < y_0$, par conséquent $u > 0$ pour $t < 0$.

Conidérons à présent le cas $t > 0$. Si $t \to T_+$ alors $x \to +\infty$, car sinon x serait borné et $y \to +\infty$, $\dot{x} \to +\infty$, contradiction avec le lemme 3. Si y est borné $u \to +\infty$. Sinon $\dot{u} = xy + ax^2 + 1 - \alpha x$, $\dot{u} \to +\infty$ car y est croisant, et comme $\dot{y} = u + \alpha y + \beta$ le lemme 3 où on a fait le changement de variables $X = u$ montre que $u \to +\infty$. Donc la droite ne redresse pas le champ.

Par conséquent aucune droite ne redresse le champ.

On considère la courbe Γ

$$\Gamma : y = \frac{1-ax^2}{x} \quad x > 0$$

et le domaine D dont le bord est Γ

$$D = \{(x,y) | \quad xy + ax^2 - 1 > 0, \quad x > 0\}.$$

La courbe Γ est transverse au champ, car la dérivée de la quantité $y + ax - 1/x$, lorsque cette quantité est nulle est $x + 2a + 2x^{-2} > 0$ si $x > 0$.

Montrons que Γ redresse le champ. Soit P le demi-plan $x < 0$ et D_1 le complémentaire de $D \cup P$.

Soit (x_0, y_0) un point. S'il est dans D, alors $\dot{x} > 0$, $\dot{y} > 0$ et $D \cap \{(x,y)\ x \leq c_1, y \leq c_2\}$ est compact : la trajectoire sort de D si $t \to T_-$.

Si (x_0, y_0) est dans P la trajectoire en sort si $t \to T_+$. En effet cela est clair si $x \to +\infty$. Si x est borné alors, comme $x < 0$, on a $y \to -\infty$, donc $\dot{x} \to +\infty$, ce qui contredit le lemme 3. Enfin si $x \to -\infty$, $\dot{y} \to -\infty$; dans le cas $a \neq 0$, on a $\dot{x} = x(ax+y) + 1 \to +\infty$ (car y est majoré), ce qui est absurde; sinon, par le lemme 3 où on a échangé les rôles de x et y, $y \to -\infty$, donc $\dot{x}$ devient > 0 ce qui est également absurde.

Donc la trajectoire rentre dans D_1. Considérons une condition initiale dans D_1. La trajectoire ne peut pas rentrer dans le demi-plan $x < 0$, puiisque $\dot{x} = 1$ si $x = 0$, et x ne peut tendre vers 0, car sinon y tendrait vers $+\infty$ et $\dot{x}$ deviendrait > 1. Il existe donc k avec $x \geq k > 0$ (lemme 2). Dans D_1, y est croissant et $(D_1 \cup \Gamma) \cap \{(x,y) | x \geq k, y \geq c\}$ est compact si $a \neq 0$, et dans ce cas la trajectoire, qui ne peut rester dans ce compact, coupe Γ et rentre dans D. Si $a = 0$, y devient positif, car sinon y est borné, $x \to \infty$, $\dot{y} \to \infty$ ce qui contredit le lemme 3. On conclut comme précédemment en remarquant que $(D_1 \cup \Gamma) \cap \{(x,y) | x \geq k, y \geq c\}$ est compact si $c > 0$.

12.3. Etude de CP2.c

$$\begin{cases} \dot{x} = (ax+y)(bx+y) + 1 \\ \dot{y} = x + c \end{cases} \qquad (CP2c)$$

Comme il n'y a pas de points fixes c'est que $(b-a)^2 c^2 < 4$.

On va découper ce cas en plusieurs sous-cas.

12.3.1. Cas $ab < 0$

Le champ ne se redresse pas.

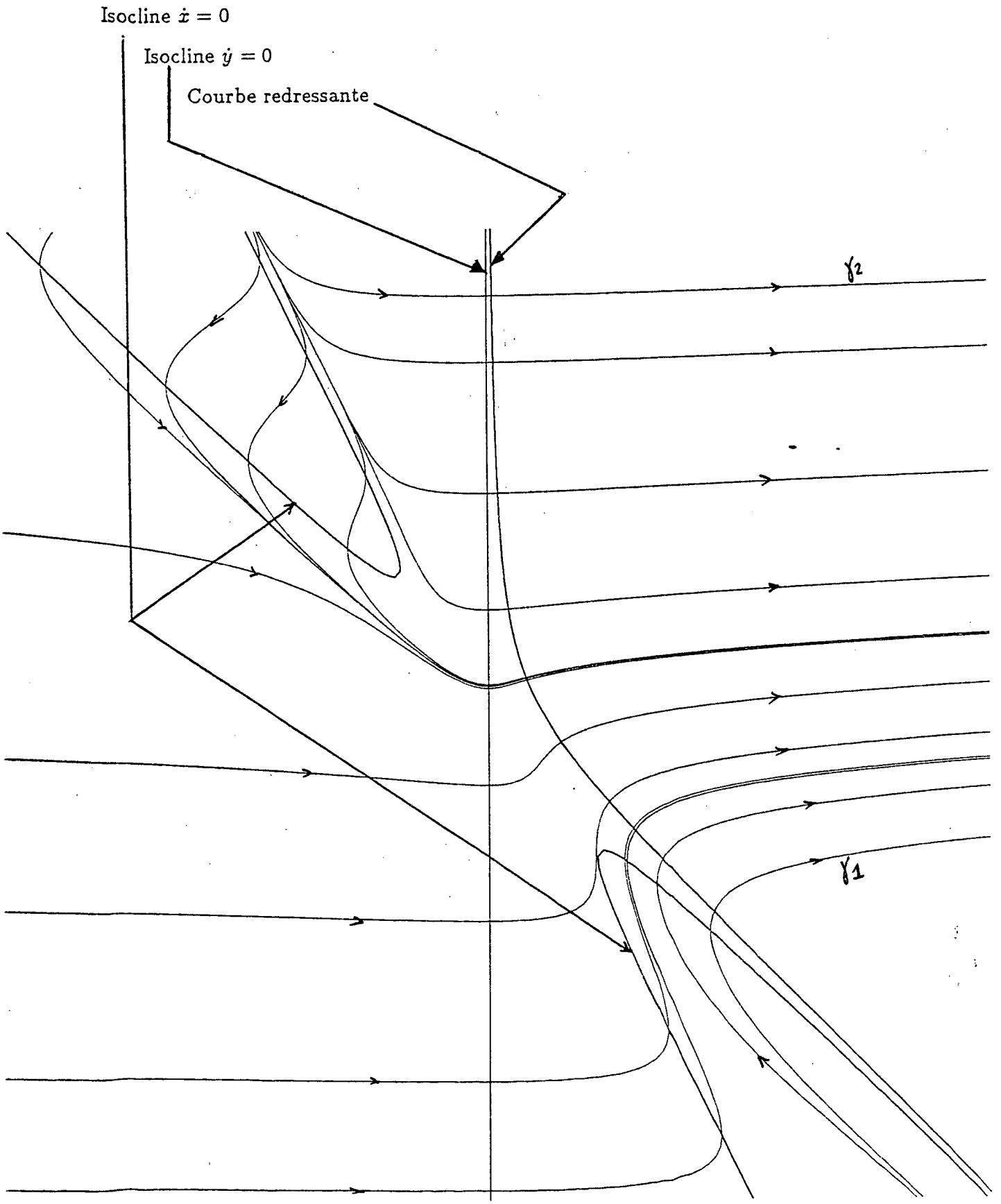
On prend $\gamma > 0$ avec $-\delta = \gamma^2 + \gamma|a+b| + ab < 0$. On choisit $A < 0$ de sorte que, si $x \leq A$, on ait $x + c < 0$, $1 - \delta x^2 < 0$ et $x + c - \gamma + \gamma\delta x^2 > 0$.

Soit $w = y - \gamma x$. On suppose initialement $w \geq 0$, $x \leq A$ et $y \leq 0$ (il suffit de prendre $x = A$ et $y = 0$).

Alors, tant que ces conditions sont vérifiées, on a

Cas CP2c

$$\begin{cases} \dot{x} = (x+y)(2x+y)+1 \\ \dot{y} = x + 1/4 \end{cases}$$



$\dot{x} = y^2 + xy(a+b) + abx^2 + 1 \leq y^2 + |xy(a+b)| + abx^2 + 1 \leq x^2(\gamma^2 + \gamma|a+b| + ab) + 1 = 1 - \delta x^2 < 0$

$\dot{w} = x + c - \gamma\dot{x} \geq x + c - \gamma(1 - \delta x^2) = x + c - \gamma + \gamma\delta x^2 > 0$

$\dot{y} = x + c < 0.$

Par conséquent les hypothèses restent toujours vérifiées le long de la trajectoire positive issue de (x, y).

Donc pour $t > 0$ on a $x < A$. Regardons ce qui se passe si $t \to T_-$. On ne peut pas couper la droite $x = -c$, car sur cette droite on a $\dot{x} > 0$.

On a donc trouvé une trajectoire sur laquelle $x < -c$.

Changeant x en $-x$ et t en $-t$, donne le même système, où a, b et c sont remplacés par leurs opposés. Il existe donc une trajectoire avec $x > c$. Comme par ailleurs $\dot{y} = 0$ est un minimum absolu en y, on conclut comme dans le cas CP2b avec $a < 0$, en construisant un piège compris enntre ces deux trajectoires et une hypothétique courbe redressante.

12.3.2. Cas $a = b$

Ce champ se redresse par le théorème 2.

12.3.3. Cas $ab > 0$ $a \neq b$

Ce champ se redresse mais pas par des droites.

Quitte à faire $X = -x$, $T = -t$ on peut supposer $a > 0$ et $b > 0$.

On pose $u = ax + y$, $v = bx + y$. Alors

$$\begin{cases} \dot{u} = a[u - \dfrac{1}{a(a-b)}][v + \dfrac{1}{a(a-b)}] + a + c + \dfrac{1}{a(a-b)^2} \\ \dot{v} = b[u - \dfrac{1}{b(a-b)}][v + \dfrac{1}{b(a-b)}] + b + c + \dfrac{1}{b(a-b)^2}. \end{cases}$$

En outre $\dot{u} = 0$, $\dot{v} = 0$ sont des minima absolus en u et v: montrons-le pour u. Si $\dot{u} = 0$ on a $\dot{v} = (b-a)(uv+1)$, donc $\ddot{u} = [au(b-a) + 1](uv + 1)$. De plus $v = \dfrac{u + (a-b)(a+c)}{1 + a(b-a)u}$ d'où $\ddot{u} = u^2 + (a-b)cu + 1 > 0$ car $(a-b)^2c^2 < 4$.

Montrons à présent qu'il existe une trajectoire $\gamma_1(t)$ telle que si $t \to T_+$, alors $u, v \to +\infty$ et si $t \to T_-$, v est borné et $u \to +\infty$.

On prend comme condition initiale (u_0, v_0) sur l'isocline $\dot{u} = 0$ ($v_0 = \dfrac{u_0 + (a-b)(a+c)}{1 + a(b-a)u_0}$),

avec $u_0 > 1/[b(a-b)]$, et $u+0 \neq 1/[a(a-b)]$. Alors u a un minimum absolu en $t = 0$ et la relation précédente est valable pour tout u. Regardons ce qui se passe si $t \to T_+$. Si u est borné, $|y| \to \infty$ et $|x| \to \infty$ car (u, x) et (v, y) sont des coordonnées, mais les équations donnant u et $\dot{y}$ sonnt alors contradictoires quand aux signes de x et y. Ainsi u n'est pas borné et v non plus par symétrie. Comme u est minoré, $u \to +\infty$, et d'après l'équation donnant $\dot{u}$, $v \to +\infty$ aussi nécessairement.

Regardons maintenant ce qui se passe si $t \to T_-$. Supposons que v tende vers $\epsilon\infty$, où $\epsilon = \pm 1$. D'après le choix de u_0 comme u est minoré, on aurait $\dot{v} \to \epsilon\infty$, ce qui est absurde. Par conséquent, v est borné, u tend vers l'infini, comme il est minoré, $u \to +\infty$.

Par symétrie, il existe une trajectoire $\gamma_2(t)$ vérifiant: si $t \to T_+$, $u, v \to +\infty$ et si $t \to T_-$, u est borné et $v \to +\infty$.

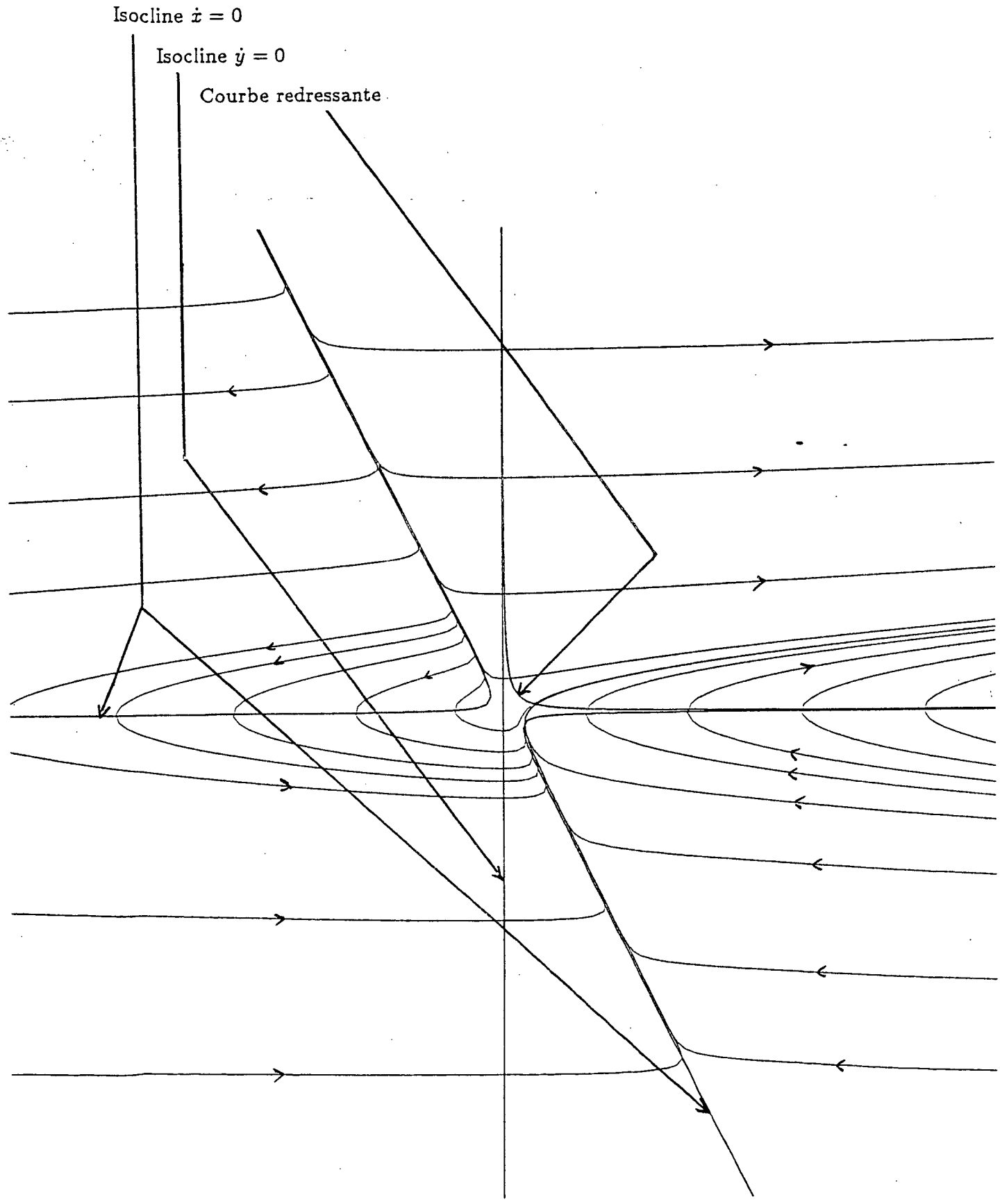
Supposons (quitte à échanger les rôles de u et v) que $a < b$. Montrons que le champ ne peut pas se redresser par des droites.

** Les droites $u = k$, $v = k$ ne redressent pas le champ car u et v ont des minima absolus.

** Les droites $w = u - \alpha v - \beta$ avec $\alpha < 0$ ne redressent pas le champ. En effet sur γ_1 on a $w \to +\infty$ si $t \to T_\pm$.

Cas CP2c

$$\begin{cases} \dot{x} = y(3x + y) + 1 \\ \dot{y} = x + 1/4 \end{cases}$$



** Les droites $w = u - \alpha v - \beta$ avec $\alpha = a/b$ ne redressent pas le champ car ce sont les droites y =Cte, qui ne sont pas transverse au champ en $x = -c$.

** Les droites $w = u - \alpha v - \beta$ avec $0 < \alpha < a/b$ ne redressent pas le champ: considérons γ_1. Si $t \to T_-$, $w \to +\infty$. Si $t \to T_+$, comme $w = y(b-a)/b + (a/b - \alpha)v - \beta$ et que $(a/b - \alpha)v - \beta \to +\infty$, on a $w \to +\infty$ aussi, car y ne peut tendre vers $-\infty$, vu que ceci entrainerait $x \to +\infty$, puisque $u \to \infty$, et donc $\dot{y} \to +\infty$.

** Les droites $w = u - \alpha v - \beta$ avec $\alpha > a/b$ ne redressent pas le champ. En effet, considérons γ_2. Si $t \to T_-$ alors $w \to -\infty$. Si $t \to T_+$, $\dot{w} \to -\infty$ car $\dot{w} = (1-\alpha)[(u-v)/(a-b)+c] + (a - \alpha b)(uv+1)$. Comme $\dot{y} = [w + (\alpha - 1)y + \beta]/(a - \alpha b) + c$, le lemme 3 s'applique aux variables (w, y) et montre que $w \to -\infty$.

En conclusion, le champ ne se redresse pas par une droite. Il se redresse comme on va le voir par une autre courbe.

On revient aux variables initiales. On va traiter en même temps le cas $ab = 0$. On suppose $0 \leq a < b$.

On considère la courbe Γ

$$\Gamma : y = \frac{1 - ax(x+c)}{x+c} \quad x + c > 0$$

et le domaine D dont le bord est Γ

$$D = \{(x,y) | \quad (ax+y)(x+c) \geq 1, \quad x + c > 0\}.$$

Sur D on a $\dot{x} > 0$ et $\dot{y} > 0$. Ceci est évident pour $\dot{y}$. Posant $z = y + ax$ on a $\dot{x} = z(b-a)x + z^2 + 1$. On sait que $z(b-a) > 0$, et l'expression est linéaire en x. Sa valeur minimale à z fixé est pour $x = -c$. Pour cette valeur de x, le discriminant en z est $(b-a)^2c^2 - 4 < 0$, d'où la conclusion pour $\dot{x}$.

Si $w = (y + ax)(x + c) - 1$, on a le long d'une trajectoire $\dot{w} = (x+c)^2 + [a(x+c) + y + ax]\dot{x} > 0$, et donc Γ est transverse au champ.

Comme $D \cap \{(x,y) | \ x \leq k_1, y \leq k_2\}$ est compact, toute trajectoire de D sort de D, et coupe Γ si $t \to T_-$.

Considérons maintenant une condition initiale vérifiant $x + c < 0$. Faisons tendre t vers T_+, et montrons par l'absurde que x devient $> -c$. Sinon, y est toujours décroissant. S'il était borné on aurait $x \to -\infty$; si $ab \neq 0$, on en déduit $\dot{x} \to +\infty$, absurde; si $ab = 0$, comme $\dot{y} \to -\infty$, ceci contredit le lemme 3 où les rôles de x et y sont échangés. Si y est non borné, $y \to -\infty$, $\dot{x} \to +\infty$, $x \to +\infty$ par le lemme 3, ce qui est absurde.

Considérons maintenant une condition initiale vérifiant $x > -c$. Notons que si $t \to T_+$, x reste supérieur à $-c$ car si $x = -c$, on a $\dot{x} > 0$. Donc $\dot{y} > 0$, et y est minoré. Si x était borné, on aurait $y \to +\infty$, $\dot{x} \to +\infty$, ce qui contredit le lemme 3. Donc $x \to +\infty$. Si $ab \neq 0$, la trajectoire rentre dans D, mais si $ab = 0$, comme $\dot{y} \to +\infty$, on sait par le lemme 3 que $y \to +\infty$, donc la trajectoire rentre aussi dans D.

Par conséquent Γ redresse le champ.

12.3.4. Cas $ab = 0$

Le champ se redresse mais pas par des droites.

Comme précédemment on peut supposer $a \geq 0$ et $b \geq 0$. Par symétrie on sait se ramener à $a = 0$, $b > 0$.

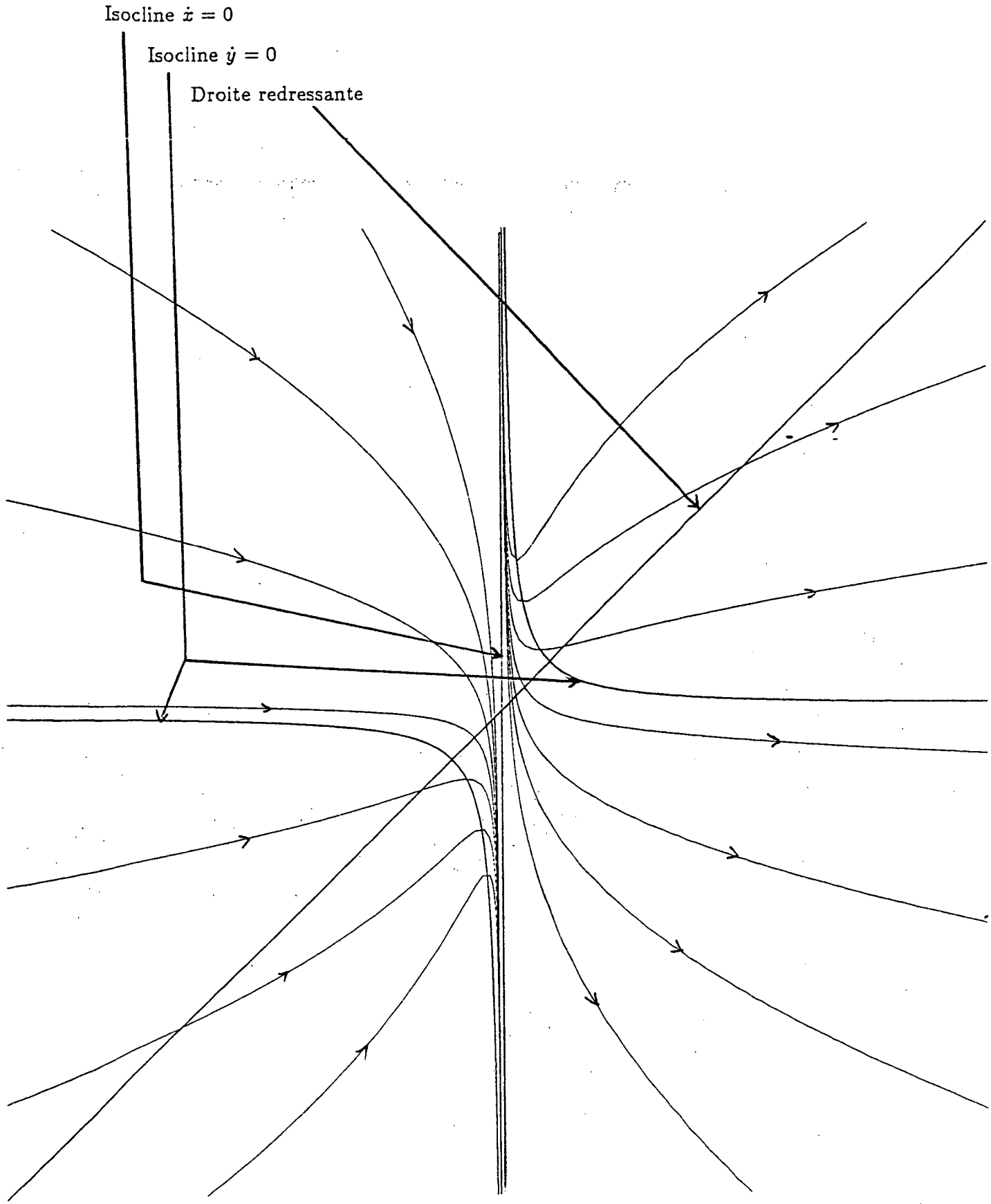
Soit δ avec $b\delta + 1 < 0$. On pose $\gamma = \delta - 1/b$ et donc $b\gamma + 2 < 0$.

On considère une trajectoire passant par la condition initiale (x_0, y_0) sur l'isocline $\dot{x} = 0$, avec $y_0 < 0$ et $x_0 = -\dfrac{y_0^2 + 1}{by_0}$. On choisit y_0 suffisamment proche de 0 pour que, si $x \geq x_0$, on ait $x > |c|$, et $\dfrac{1 + b\gamma}{\gamma^2}x^2 + 1 - \gamma x - \gamma c < 0$, (d'où $\dfrac{1 + b\gamma}{\gamma^2}x^2 + 1 < 0$). De plus on suppose que $x_0 - \gamma y_0 > 0$, et $2y_0 + bx_0 > 0$. On pose $v = x - \gamma y$.

Pour $t = 0$, x a un minimum local et $v > 0$. Il existe donc un plus petit nombre $t_1 < 0$ tel que sur $]t_1, 0[$ on ait $\dot{x} < 0$ et $v > 0$. Montrons par l'absurde que $t_1 = T_-$. En effet, $v(t_1) = 0$ est absurde, car $x \geq x_0$ et la deuxième condition imposée à x_0 s'écrit $\dot{v} < 0$ si $v = 0$. Donc $\dot{v}$ a le mauvais signe. L'hypothèse $\dot{x} = 0$ est

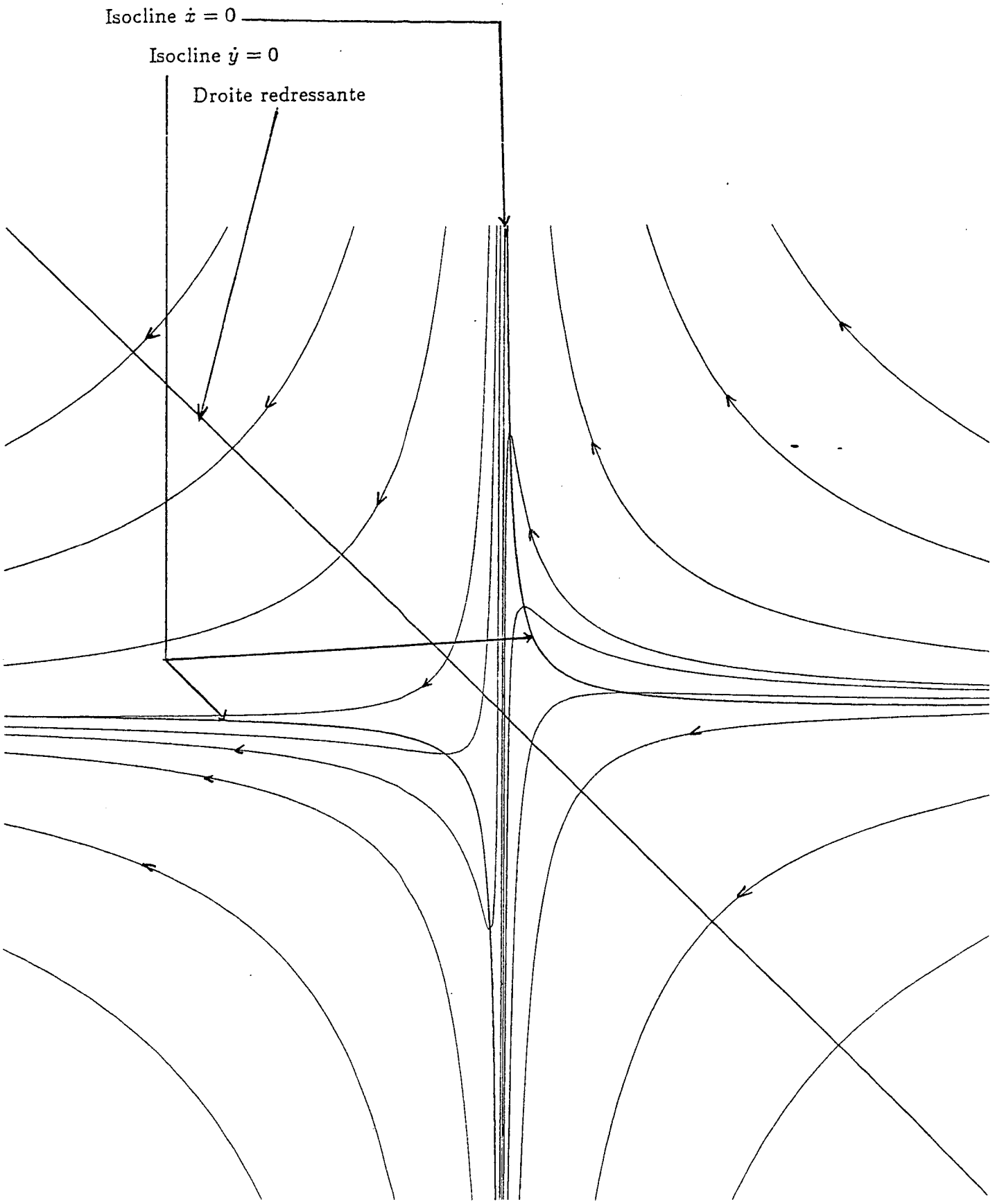
Cas CP0

$$\begin{cases} \dot{x} = 2x^2 \\ \dot{y} = xy - 1 \end{cases}$$



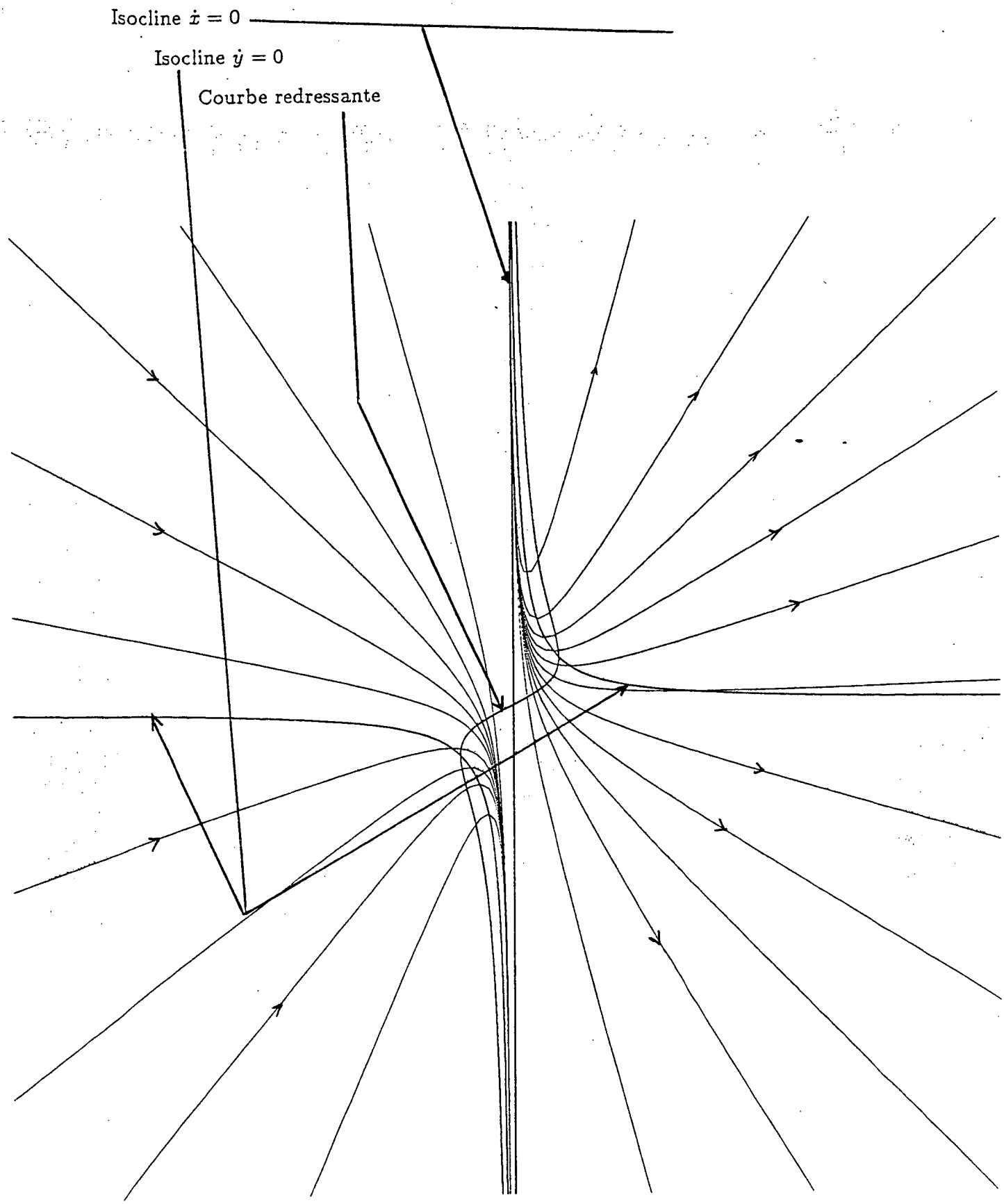
Cas CP0

$$\begin{cases} \dot{x} = -x^2 \\ \dot{y} = xy - 1 \end{cases}$$



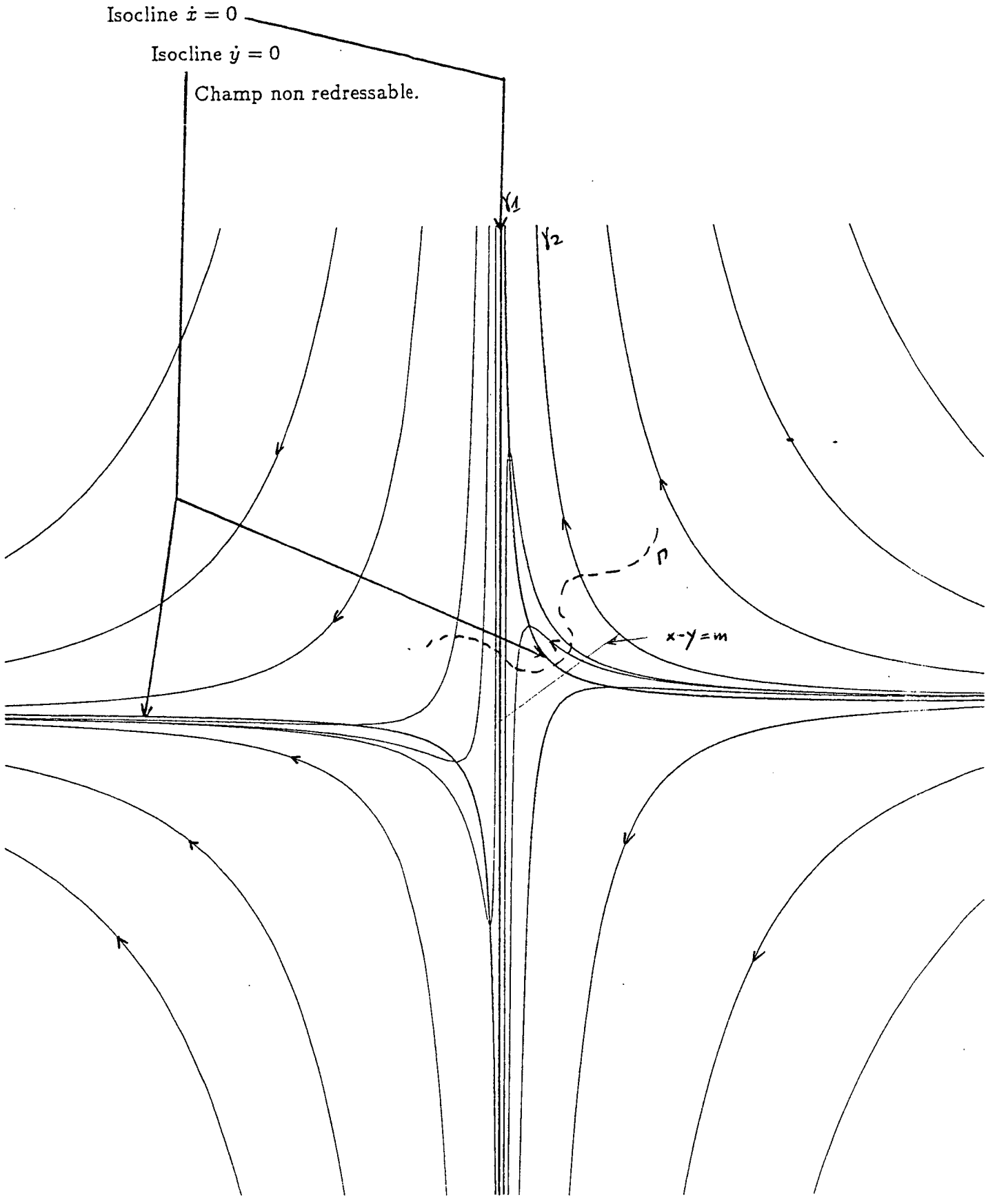
Cas CP0

$$\begin{cases} \dot{x} = x^2 \\ \dot{y} = xy - 1 \end{cases}$$



Cas CP0

$$\begin{cases} \dot{x} = -x^2/2 \\ \dot{y} = xy - 1 \end{cases}$$



également absurde, car si $\dot{x} = 0$ on a $\ddot{x} = b\dot{y}v + \dot{y}(b\gamma + 2)y > 0$, car $\dot{y} > 0$, $y < 0$ et $v > 0$, et $\ddot{x}$ serait du mauvais signe.

Par application du lemme 3, on en déduit que, si $t \to T_-$, ni y ni v ne peuvent être bornés, et par conséquent $v \to +\infty$, $y \to -\infty$, d'où $x \to +\infty$ et $x - \delta y \to +\infty$.

Considérons maintenant une droite $u = x - \alpha y - \beta$. Si $b\alpha + 1 < 0$ on prend $\delta = \alpha$ et les calculs précédents montrent que si $t \to T_-$, $u \to +\infty$. Si $b\alpha + 1 \geq 0$, on prend $\epsilon > 0$, $\delta = -1/b - \epsilon$ d'où $u = v - y(\alpha + 2/b + \epsilon) - \beta \to +\infty$.

Faisons maintenant tendre t vers T_+ sur la trajectoire en question. Initialement x et y sont croissants. Tant que x est croissant on a $x > -c$, et donc $\dot{y}$ ne peut s'annuler. Si $\dot{x}$ s'annulait, on aurait $\ddot{x} = \dot{y}(2y + bx)$. Or $\dot{y} > 0$, et $2y + bx$ était initialement positif et est croissant, donc $\ddot{x}$ serait du mauvais signe. Donc pour $t > 0$, x et y croissent. Si y est borné on a $x \to +\infty$, d'où $u \to +\infty$, sinon $y \to +\infty$, $\dot{x} \to \infty$ et par le lemme 3, $x \to +\infty$. Si $\alpha \leq 0$ on en déduit $u \to +\infty$. Supposons maintenant $\alpha > 0$. On peut supposer $u_0 > -\beta$. Alors si $u = -\beta$ on a $\dot{u} = (1 + b\alpha)\alpha^{-2}x^2 + 1 - \alpha(x + c)$. On peut choisir x_0 tel que si $x \geq x_0$ on ait $\dot{u} > 0$ pour $u = -\beta$, ce qui signifie que, pour tout $t > 0$, $u > -\beta$. Donc u ne peut tendre vers $-\infty$. Si u était borné, on aurait $\dot{u} \to +\infty$, et par le lemme 3, $u \to +\infty$. Donc dans tous les cas $u \to +\infty$.

La droite ne redresse donc pas le champ.

Les droites $y = k$ ne redressent pas non plus, car ne sont pas transverses au champ.

Par conséquent aucune droite ne redresse le champ.

On a vu cependant dans la sous-section précédente que le champ était redressable par une courbe Γ.

13. Etude de CP0

$$\begin{cases} \dot{x} = ax^2 \\ \dot{y} = xy - 1 \end{cases} \qquad (CP0)$$

On va distinguer plusieurs cas en fonction du signe de a et intégrer explicitement le champ.

Supposons que l'instant initial soit $t = 0$. Lorsque $x_0 \neq 0$ posons $v = k - at$ où $k = 1/x_0$ de sorte que $x = 1/v$. Les trajectoires sont constituées de $x = 0$, $y = -t + Cte$, des trajectoires avec $v > 0$ et des trajectoires avec $v < 0$. Comme le champ est pair il suffit d'étudier les trajectoires avec $v > 0$.

** Cas $a = -1$.

Ici $x = 1/v$, $y = v(C - \log v)$.

Le champ se redresse par $x + y = 0$, car $x + y = \dfrac{1}{v}(1 + v^2C - v^2 \log v)$.

Si $v \to 0$, $x + y \to +\infty$ et si $v \to +\infty$, $x + y \to -\infty$. De plus $\dot{x} + \dot{y} = x(y - x) - 1 = -2x^2 - 1 < 0$ si $x + y = 0$.

Si $a \neq -1$ on a $x = \dfrac{1}{v}$, $y = \dfrac{v}{a+1} + \dfrac{C}{v^{1/a}}$.

** Si $a > 1$, $x - y = \dfrac{1}{v}(1 - \dfrac{v^2}{a+1} - Cv^{1-1/a})$

donc si $v \to 0$, $x - y \to +\infty$ et si $v \to +\infty$, $x - y \to -\infty$.

Donc $x = y$ redresse le champ, car si $x = y$ on a $\dot{x} - \dot{y} = x(ax - y) + 1 = x^2(a - 1) + 1 > 0$.

** Si $a < -1$, $x + y = \dfrac{1}{v}(1 + \dfrac{v^2}{a+1} + Cv^{1-1/a})$

donc si $v \to 0$ $x + y \to +\infty$ et si $v \to +\infty$ $x + y \to -\infty$.

Donc $x + y = 0$ redresse le champ, car si $x + y = 0$ on a $\dot{x} + \dot{y} = x(ax + y) - 1 = x^2(a - 1) - 1 < 0$.

** Si $-1 < a < +1$, le champ ne se redresse pas par des droites, car si $u = y - \alpha x - \beta$ on a

$u = \frac{1}{v}(\frac{v^2}{a+1} + Cv^{1-1/a} - \alpha - \beta v)$.

Si $a > 0$, on prend $C > 0$ alors $u \to +\infty$ si $v \to 0$ ou $v \to +\infty$. Si $a < 0$ on prend C du signe de $-\alpha$ alors u tend vers l'infini avec le signe de C si v tend vers 0 ou l'infini.

** Si $0 < a \leq 1$ on pose $w = x - \frac{2y}{y^2+1}$. Si $w = 0$ on a $\dot{w} = \frac{2[(2a+1)y^4 + (2a-2)y^2 + 1]}{(y^2+1)^3}$. Le discriminant du numérateur est $a(a-4) < 0$. Le champ est donc transverse à la courbe Γ

$$\Gamma : x = \frac{2y}{y^2+1}.$$

La courbe $x = 0$ coupe Γ. Pour les courbes avec $x > 0$ on a, si $x \to +\infty$, dans le cas $C \neq 0$, $y \sim Cx^{1/a}$, donc $y \to \infty$, donc $w \sim x$. Si $C = 0$, $y = 1/[x(a+1)]$, donc $y \to 0$ et de même $w \sim x$. Si $x \to 0$, $y \to \infty$, $y \sim [x(a+1)]^{-1}$ et donc $w \sim -(2a+1)x < 0$. La courbe Γ redresse le champ.

** Cas $-1 < a < 0$

Le champ ne se redresse pas. On prend la trajectoire γ_1 correspondant à $x = 0$, et la courbe γ_2 correspondant à $C = 0$. S'il existait une courbe redressante Γ, elle couperait ces deux trajectoires. Sur la portion de Γ entre ces deux trajectoires, $x - y$ prend une valeur maximale, disons $m - 1$. Il suffit de montrer qu'il existe une trajectoire γ_3 vérifiant $x_3 > 0$, et $y_3 < y_2$, telle que $x_3 - y_3 \geq m$, où les indices 2 et 3 correspondent aux trajectoires γ_2 et γ_3. Or, posant $u = x - y$, on a $\dot{u} = -xy + ax^2 + 1$ et si $\dot{u} = 0$, $\ddot{u} = -ax[(1-a)x^2 + 1]$, ce qui signifie que, pour $x > 0$, les extrema de u sont des minima. Il suffit alors de vérifier que les équations $\dot{u} = 0$ et $u = m$ ont bien une solution en x et y vérifiant $x > 0$, ce qui est clair, car écrivant y en fonction de x donne l'équation du second degré $(a-1)x^2 + mx + 1 = 0$, qui a deux racines de signes opposés. On vérifie alors que pour cette valeur de x, on a $y = x - m$ est bien $< 1/[x(a+1)]$ ce qui montre que la trajectoire trouvée est en dessous de γ_2.

Conclusion de ce cas :

* si $a \leq -1$ le champ se redresse par une droite
* si $-1 < a < 0$ le champ ne se redresse pas
* si $0 < a \leq 1$ le champ se redresse mais pas par une droite
* si $a > 1$ le champ se redresse par une droite.

14. Bibliographie

[1] W.A. Coppel *A survey of quadratic systems*, Journal of Diff. Eq. 2, 293-304 (1966)

[2] M.P. Muller *Quelques propriétés des feuilletages polynômiaux du plan*, Bol. Soc. Math. Mex. (2) 21 no 1, 6-14 (1976)

[3] S. Schecter & M.F. Singer *Planar polynomial foliations*, Proc. of Amer. Math. Soc. 79, 4 (1980)

[4] A.N. Berlinski *The topology of a family of integral curves of a homogeneous differential equation*, Diff. Eq. 8 (1972) 3, 297-304

[5] L.S. Lyagina *The integral curves of the equation* $y' = \frac{ax^2+bxy+cy^2}{dx^2+exy+fy^2}$, Uspehir. Mat. Nauk. 6 (1951) no 2 (42)

[6] V. Arnold, Equations différentielles ordinaires, ed. Mir, (Moscou) 1974

[7] Coddington & Levinson, Theory of ordinary differential equations, McGraw Hill, 1955.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique